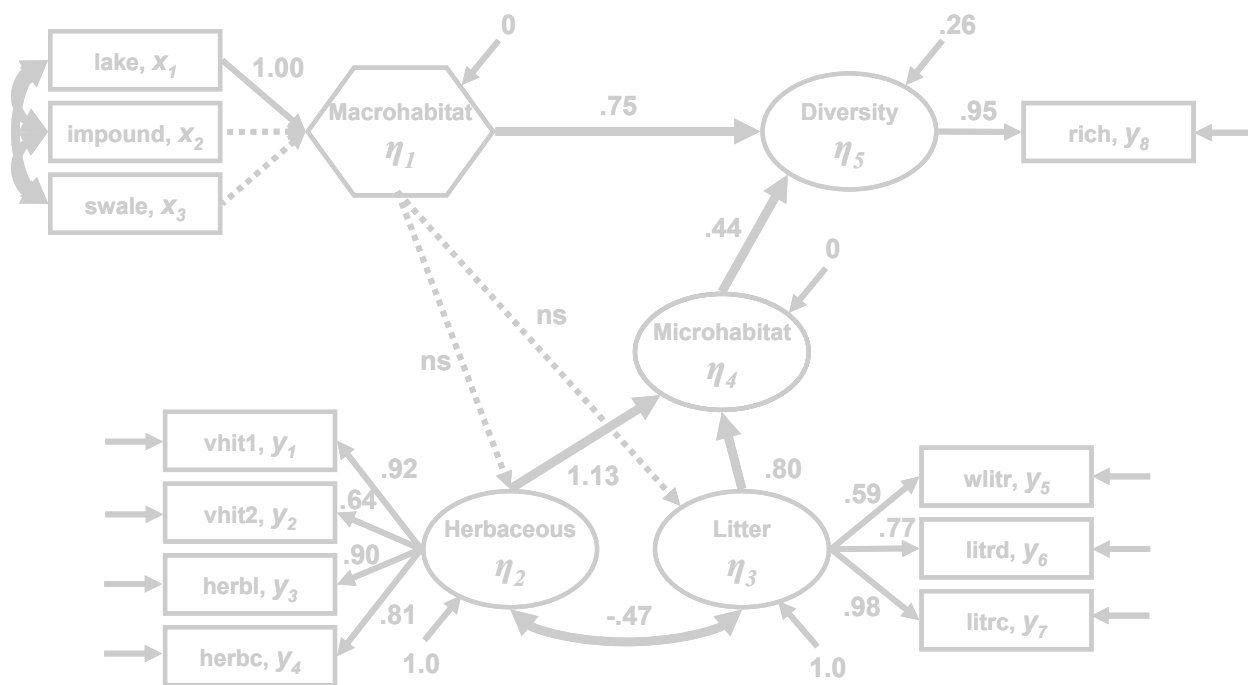


The Interface Between Theory and Data in Structural Equation Models

By James B. Grace and Kenneth A. Bollen



Open-File Report 2006-1363

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia 2006

For product and ordering information:
World Wide Web: <http://www.usgs.gov/pubprod>
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth,
its natural and living resources, natural hazards, and the environment:
World Wide Web: <http://www.usgs.gov>
Telephone: 1-888-ASK-USGS

Suggested citation:
Grace, J.B., and Bollen, K.A., 2006, The interface between theory and data in structural
equation models: U. S. Geological Survey Open-File Report 2006-1363, 33 p.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply
endorsement by the U.S. Government.

Although this report is in the public domain, permission must be secured from the individual
copyright owners to reproduce any copyrighted material contained within this report.

The Interface Between Theory and Data in Structural Equation Models

By James B. Grace¹ and Kenneth A. Bollen²

Abstract

Structural equation modeling (SEM) holds the promise of providing natural scientists the capacity to evaluate complex multivariate hypotheses about ecological systems. Building on its predecessors, path analysis and factor analysis, SEM allows for the incorporation of both observed and unobserved (latent) variables into theoretically based probabilistic models. In this paper we discuss the interface between theory and data in SEM and the use of an additional variable type, the composite, for representing general concepts. In simple terms, composite variables specify the influences of collections of other variables and can be helpful in modeling general relationships of the sort commonly of interest to ecologists. While long recognized as a potentially important element of SEM, composite variables have received very limited use, in part because of a lack of theoretical consideration, but also because of difficulties that arise in parameter estimation when using conventional solution procedures. In this paper we present a framework for discussing composites and demonstrate how the use of partially reduced form models can help to overcome some of the parameter estimation and evaluation problems associated with models containing composites. Diagnostic procedures for evaluating the most appropriate and effective use of composites are illustrated with an example from the ecological literature. It is argued that an ability to incorporate composite variables into structural equation models may be particularly valuable in the study of natural systems, where concepts are frequently multifaceted and the influences of suites of variables are often of interest.

Introduction

The application of path analysis to problems in the natural sciences has a long history (Wright, 1921, 1968). Only recently has the capacity for evaluating multivariate hypotheses using the more comprehensive approach known as structural equation modeling (SEM) attracted the interest of ecologists (Johnson and others, 1991; Mitchell, 1992; Pugeseck and Tomer, 1996; Shipley, 2000; Pugeseck and others, 2003; Grace 2006). At present, there appears to be a rapidly growing interest in SEM in the natural sciences. Reasons that SEM may prove to be of utility to researchers studying natural systems include its capacity for representing and testing multivariate hypotheses and its flexibility as a statistical modeling framework. Several of the recent suggestions for how to advance ecological sciences given by Belovsky and others (2004), such as a better integration of theory and empirical evidence and the need to study multiple causes simultaneously, are enhanced by SEM. It can be argued that the traditional univariate hypothesis testing framework limits our ability to study interacting systems and that SEM may provide some relief from this limitation. Along with the potential utility of SEM, however, are certain challenges to matching models to theories.

A key feature of SEM is the capacity to include latent (unobserved) variables along with observed ones in models. The development of an appropriate structural equation model (SE model) is driven by the nature of the theoretical constructs or ideas that represent the mechanisms of interest, as well as the variables measured and the nature of the data. Often, general concepts are represented in SE models by using latent variables. In such models, latent variables are often used to represent underlying

¹ U.S. Geological Survey, National Wetlands Research Center, 700 Cajundome Blvd., Lafayette, LA 70506.

² Department of Sociology, CB 3210 Hamilton Hall, University of North Carolina, Chapel Hill, NC 27599.

causes, while observed variables represent the effects (manifestations) of the latent factors. However, a growing body of literature indicates that in many cases, constructs are more appropriately modeled in a fundamentally different way, with observed variables representing a collection of causes rather than manifestations of effects. Frequently, when constructs and the available indicators (indicators are observed variables related to a construct) possess this sort of relationship, the constructs are more appropriately modeled by using composite variables, which differ in their properties and interpretations from conventional latent variables.

As early as 1964, Blalock pointed out that the classic representation of latent variables influencing their indicators is not always appropriate. In particular, indicator variables might be determinants of the construct rather than effects of it. Blalock (1964) referred to the former, classic indicators as “effect” indicators since the indicators are effects of the latent variable. The latter indicators are “causal” indicators since these indicators are modeled as causes instead of effects. Periodically, practitioners of SEM have continued to state the need for careful consideration of the best way to represent theoretical constructs in models. Bollen (1984) cautioned that effect indicators should have “internal consistency” (i.e., they depend on a single entity and, therefore, would be expected to be correlated and joint responsive) while causal indicators need not.

A number of authors have continued to discuss the specification of constructs in SE models (Bollen and Lennox, 1991; MacCallum and Browne, 1993), and there has been an increased interest in the subject emerging in recent years (Bollen and Ting, 2000; Diamantopoulous and Winklhofer, 2001; Edwards, 2001; Fayers and Hand 2002; Jarvis and others, 2003; Williams and others, 2003). This literature has provided a number of useful insights, diagnostic procedures, and approaches to the question of how to treat causal indicators. A related issue is the treatment of constructs composed entirely by causal indicators, which may invoke the researcher to use composite variables to represent relationships. A clear, comprehensive, and satisfying treatment of composites has just begun to emerge (Grace and Bollen in press). In this report we expand on the main issues associated with using composites in models discussed by Grace and Bollen (in press) and provide both diagnostic and solution procedures. Further, we use an ecological example to demonstrate how composites can be used to represent constructs in SE models that match up with general theories.

Framework and Terminology

A consideration of the subject of composites requires that we have an adequate framework and sufficient terminology to express clearly the relevant issues. The theoretical background for the analysis plays an important part in how we proceed. It is typical in SEM applications that the investigators wish to evaluate some a priori hypotheses. Indeed, one of the dominant features of SEM is its capacity for leading to the rejection of multivariate hypotheses based on comparisons between data and model expectations (e.g., using chi-square goodness of fit tests based on likelihood ratios). It is also characteristic of SEM practice that the analysis is based on evaluating an a priori model rather than a null hypothesis. Thus, we expect to give a high priority to theoretical knowledge in both the processes of constructing and interpreting our SE models.

One way to begin the modeling process is by defining the relevant theoretical concepts (also known as “constructs”). The meaning that a person assigns to a construct is given in its theoretical definition. A theoretical definition explains in as simple and precise terms as possible the meaning of a construct (Bollen, 1989, p. 180), as well as the number of distinct components or aspects of the construct. The “construct model” represents the hypothesized relationships among constructs based on known or expected underlying mechanisms. Figure 1 provides an overall framework for representing the relationships between general theories, construct models, and SE models. The construct model represents the researchers’ theoretical ideas in a form that facilitates the construction of the specific SE model(s) by expressing the general, expected statistical associations. Thus, the construct model can be seen as a useful bridge between the abstractions of theory and the realities of data modeling. In the development of SE models that are meant to relate to general theories, there is the need to consider carefully the relationships between the available data and the constructs of interest. This paper is largely about these relationships between construct models and SE models and the roles that composite variables might play.

The specification of SE models has typically involved the use of both observed and latent variables. Here we define a “latent variable” as one that represents a factor which we believe to exist and be relevant to our analysis but which is unmeasured (Bollen, 2002). Latent variables are often used in SE models to represent such things as the true value of a

parameter, an underlying cause, or an important piece of missing (unmeasured) information. Latent variables play a special role in SEM because they represent a kind of bridge between observed data and theoretical generalization. As an example of how latent variables might be used, if a construct is envisioned to have distinct dimensions, then each dimension might be represented by a single latent variable. In such a case, a two-dimensional construct would be represented by two latent variables in the SE model.

“Composites” are another type of variable that can be specified in SE models, and they represent collections of causes. Composites can reflect the effects of sets of either manifest or latent variables. A composite can, thus, be referred to as either a “composite of manifest variables” or a “composite of latent variables.”³ We also recognize that the issue of how to represent a construct deals with the interrelations between a latent or composite variable and the associated variables that give it meaning. To aid our discussion, we define a “block” as the basic unit to consider in an SE model and illustrate several types of blocks in figures 2 and 3. The reader should also be aware that we use the terms “observed” and “manifest” interchangeably to refer to the directly measured variables. The term “indicator” refers to an observed or manifest variable that is linked to either a latent or composite variable.

In Figure 2 we see that given a set of three manifest variables, $x_1 - x_3$, there are various ways they can be interpreted (and thus represented) in an SE model depending on how the data relate to the theoretical construct. The most common representation is the $L \rightarrow M$ block, where directed paths run from latent to manifest variables. In this block type, the latent variable is postulated to cause, at least in part, the correlations/covariances among the manifest variables. A great deal has been written about the properties of $L \rightarrow M$ blocks and they form the backbone of latent variable modeling (Bollen, 1989). Some of the expected properties of blocks of this type will be considered later in the paper, along with diagnostics that help evaluate whether either theory or empirical relations are consistent with this representation of a construct. An alternative terminology that has been used for this case is to refer to a “latent variable with effect indicators.” We note that there is a long history of the use of the opposing terms “effect indicators” (as in the $L \rightarrow M$ block in figure 2 and “causal indicators” (as in the $M \rightarrow L$ block in figure 2) and include that usage in our discussion.

The $M \rightarrow L$ block shown in figure 2 represents the situation in which manifest variables have causal influences on a latent (unmeasured) variable. In this case, our figure implies that the latent variable does indeed exist independent from our data despite the fact that we do not have measures of it. Graphically, this property of independent existence for the latent variable is represented by the presence of an error variance ζ_l , which implies that while the x s have causal effects on η_l , they do not completely determine it.

In contrast to the $M \rightarrow L$ block is the $M \rightarrow C$ block, in which we have a composite summarizing the collective influences of $x_1 - x_3$. Since the error variance is specified to be 0 for a composite, this condition signifies that it is completely determined by its causes. We can distinguish two kinds of composites; one is the “fixed composite”. In this type, the loadings from causes are specified a priori to have particular values by way of definition. An ecologically relevant example would be the “importance value,” which is defined as the sum of the relative density, relative abundance, and relative frequency (usually for a species within a community). A second type of composite can be referred to as a “statistical composite”, which is the type considered in this paper. A statistical composite represents the collective effects of a set of causes on some response variable(s). In simplistic terms, the statistical composite is similar to a multiple regression predictor and usually represents some weighted combination of causal influences that maximizes variance explanation in one or more response variables (not shown in fig. 2).

Figure 3 presents more complex examples of blocks where relationships between latent variables are considered. The $L \rightarrow L$ block represents effects of latent causes on a latent response, where all latent variables have effect indicators. This is another standard block type found in models discussed in most introductory SEM treatments. What is represented by this block structure is that the causal mechanisms involve direct relations among latent factors and that we can understand the covariances among a set of x and y variables based on that mechanism. Also implied is that the effect indicators associated with each latent variable (e.g., $x_1 - x_3$) serve as multiple measures for that latent factor. Because the flow of causation is from the latent factors to the indicators, error terms are used to represent how well the indicators are predicted by the model, while the latent factors are presumed to be free from measurement error.

³ It is even possible to have a composite that is a mixture of observed and latent variables.

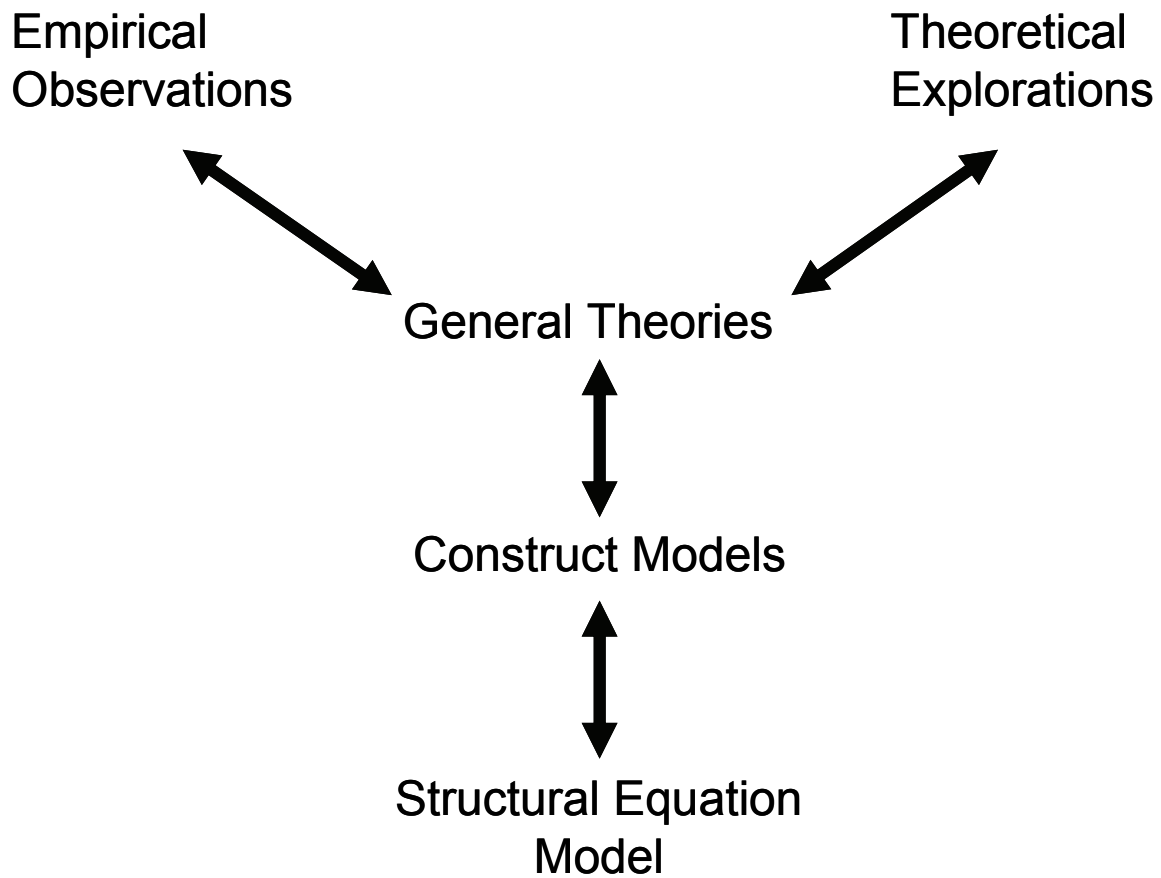


Figure 1. Overview of the relationships between general theories, construct models, and structural equation models. Note that what are referred to here as general theories are sometimes referred to as hypotheses, for example, in the case of the "intermediate disturbance hypothesis", or models, as in the "dynamic equilibrium model". Construct models refer to hypothesized relations among theoretical constructs and take a form that defines how particular structural equation models relate to general theoretical ideas. The structural equation model itself is the specific formulation of the equations used to describe relations among the included variables.

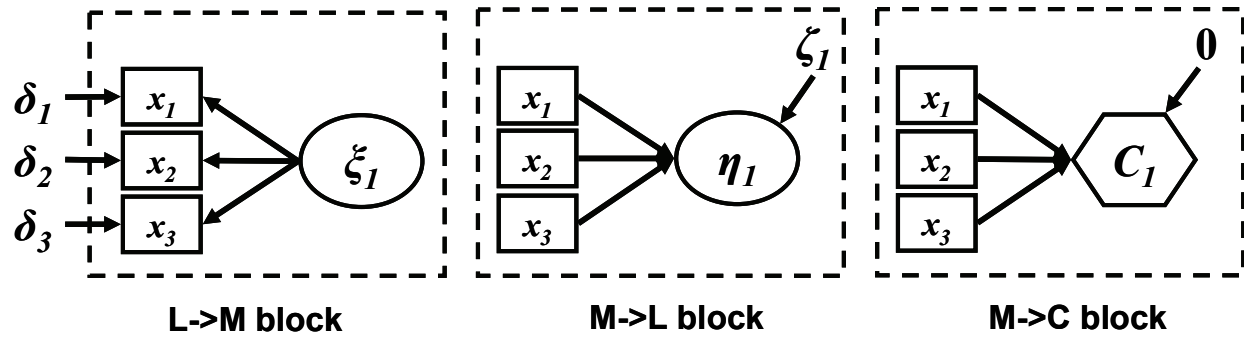


Figure 2. Representation of blocks specifying relationships between manifest (observed) indicators (in boxes) and latent variables (in circles) or composite variables (in hexagons). “L” refers to a latent variable or unmeasured cause, “M” refers to manifest variables, and “C” refers to a composite, which is a direct product of some set of causes and has zero error variance.

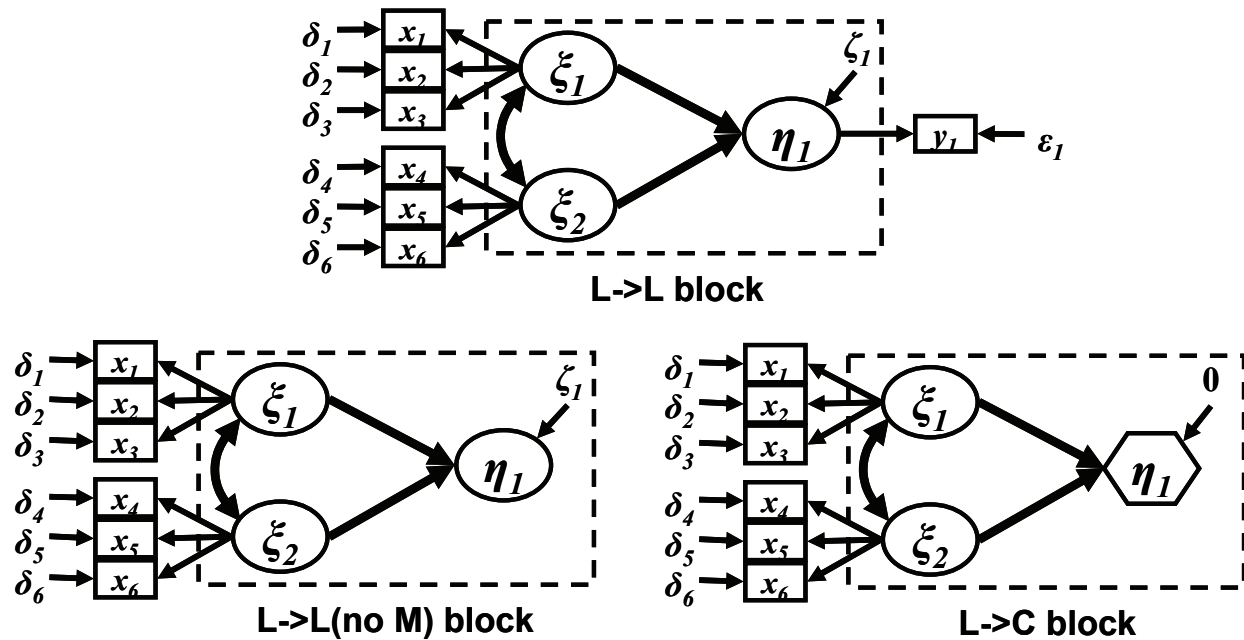


Figure 3. Representation of blocks specifying relationships between latent causes and latent or composite variables. Block terminology is as described in the caption of figure 2 except that here in this figure composites of latent variables are in circles, though still with zero error variance.

The $L \rightarrow L(\text{no } M)$ block in figure 3 represents causal influences from latent variables that have effect indicators to a latent variable for which there is no effect indicator. This block type resembles the $M \rightarrow L$ block in figure 2, except for the fact that latent causes replace the manifest ones. Similarly, the $L \rightarrow C$ block in figure 3 is analogous to the $M \rightarrow C$ block in figure 2. The relevant situations for these representations are illustrated in the example application below.

An Ecological Example: Colonization of Forest Patches by Understory Herbs

Rather than continue to elaborate on constructs, blocks, and composites in the abstract, the remaining development of ideas (as well as the development of formal notation), will be conducted in the context of a specific example (hereafter referred to as Example 1). The reader should keep in mind that while not representing an exhaustive treatment of the subject, the example chosen and procedures discussed in conjunction are broadly applicable to the general issues associated with representing constructs in SE models.

Background

In a recent paper, Verheyen and others (2003) addressed the influences of factors on the ability of herbaceous plant species to recolonize forest stands in a fragmented landscape. Their work can be viewed as an effort to evaluate a dichotomy of preexisting theories. On the one hand, island biogeographic theory suggests that successful colonization will be limited by factors such as the age of an island (in this case, a forest stand is analogous to an island) and the distance to a source population. On the other hand, the local conditions in a forest stand (such as soil conditions and the abundance of competitors) are frequently believed to be the ultimate limiting factors for successful colonization. The relative importance of effects is what is of interest in this case.

In the case of Example 1, Verheyen and others (2003) examined colonization success and related it to forest stand attributes in a nature reserve in Belgium that has had a long history of agricultural use and forest fragmentation. In this landscape, much of the original forest cover was gradually removed, and in places, discrete stands of trees were replanted,

creating a mosaic of ancient and contemporary stands. Land-use history of the 360-ha preserve was reconstructed by using maps dating back into the 1700s and aerial photographs dating back to the 1940s. Using the historical records, the age of each forest stand in the landscape was determined. Extensive surveys of forest herb species throughout the entire preserve were conducted in 1999, along with characterizations of forest understory cover and soil conditions in individual stands. Together, the sources of data allowed Verheyen and others (2003) to estimate the number of colonizations by each herb species in contemporary forest stands and the distance to the nearest population in the ancient stands, which was presumed to be a possible source for colonization. A total of 180 forest stands, with ages from 1 to 195 years old, were in the final data set. All species that were included in the analyses are obligate forest understory perennials.

Verheyen and others' (2003) initial model of constructs and the relations among them are shown in figure 4. As represented here, the authors were interested in differentiating between the effects of two exogenous (external) influences, landscape properties and soil conditions, on competitor abundance and ultimately colonization success (the latter two of which are considered endogenous or internal to the model). The authors used a two-stage analysis to evaluate the model in figure 4, first deriving composite indices, followed by a path analysis of relations among constructed indices using SEM. Here we wish to provide a more formal illustration of how their constructs and the relations to associated indicators can be evaluated. We also wish to discuss issues related to the solution of SEMs containing composites. We use a selective subset of the data analyzed by Verheyen and others (2003), and our purpose is not to confirm or contradict the biological conclusions of the original analysis but rather to illustrate statistical methods. Data characteristics were considered in this analysis, and corrective actions (e.g., transformations) were taken where needed, though the particulars are not discussed further here.

In Example 1, our starting point is the initial conceptual (construct) model summarizing the core theoretical questions posed (fig. 4). We start by considering the various ways we might address the hypotheses implied by figure 4. To address this, we must consider the available indicators and then the linkages between indicators and constructs. In this study, Verheyen and others (2003) measured eight variables in the process of developing indicators for the four constructs of interest. Two

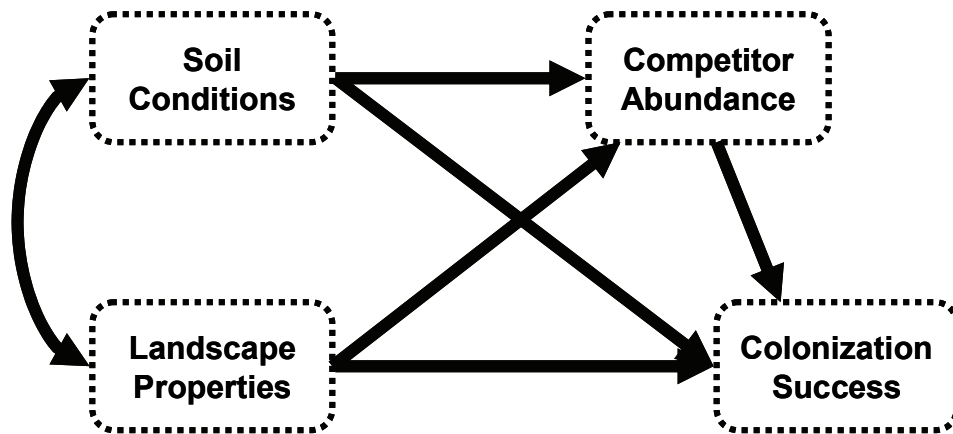


Figure 4. Initial conceptual model showing the proposed relations among constructs (in dashed rounded rectangles) relating herb colonization success to the characteristics of forest stands (from Verheyen and others 2003).

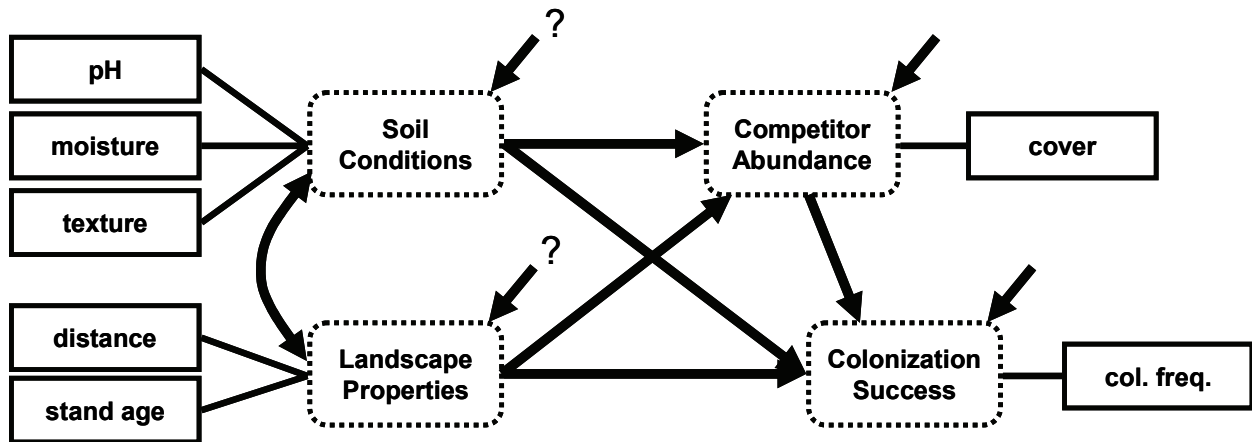


Figure 5. Associations between indicator and constructs for Example 1.

indicators of landscape conditions were measured, the estimated age of each forest stand and the distance from each reforested target stand to the nearest source patch. Three indicators of soil conditions were measured, including soil texture, soil moisture, and soil pH. Two indicators of competitor abundance were measured, the cover of herbaceous competitors and the abundance of understory shrubs. For the herb species whose colonization will be considered in this example application (*Lamium galeobdolon*), shrub abundance was not significantly related to colonization success (though herbaceous cover was). So, to simplify our example, we only use a single indicator for competitor abundance, which is herbaceous cover. The proportion of sampling points where a species was found in a forest stand serves as the single indicator of colonization success.

Figure 5 shows which measured indicators are associated with particular constructs based on the theoretical reasoning of the authors. In this initial representation, we make no attempt to express the structure of the blocks, therefore, constructs are related to observed indicators with nondirectional lines rather than directional arrows. Error terms are specified for Competitors and Colonization, while the diagram is ambiguous as to whether or not Soil and Landscape possess error terms because the directions of influences are not given.

Possible Model Structures

There are several possibilities for how the relations in figure 5 might be developed into SE models. The primary determining factor for deciding which of the possibilities is most appropriate is the nature of the causal relations. There are, in fact, a great many possibilities that could be considered, though here we will only discuss a few key ones of contrasting type. One of these possibilities is that the most appropriate model is a classic latent variable model of the sort shown in figure 6. Here we see that the block structure used to represent relations between indicators and individual latent variables is the L→M block type. In this model, which will be referred to as Model A, two of the latent variables, Competitors and Colonization, are associated with one effect indicator each. As appropriate for such relationships, arrows are directed from the causes (the latent variables) to the indicators. The two other latent variables in this model, Soil and Landscape, are associated with multiple effect indicators. Causal relations are represented among latent variables, and direct paths between manifest variables are not specified.

A second possibility is that the multiple indicators associated with Soil and Landscape are causal indicators. This type of structure is illustrated in Model B (fig. 7). Here, the arrows pointing from pH, moisture, and texture to Soil represent the presumption that the measured soil properties jointly determine the soil's influences rather than vice versa. A similar logic applies to distance and patch age, which are presumed to contribute to Landscape in this model. For the sake of simplicity, we assume in this case that the composites Soil and Landscape are completely defined by their indicators. Thus, the multi-indicator blocks are of the M→C type, and both Soil and Landscape are represented as observed composites with zero error variances.

If we assume that the causal indicators in Model B are themselves imperfect measures, which is very reasonable in this case, then we might represent the situation by using latent variables for the true values of pH, moisture, texture, distance, stand age, competitors, and colonization. Given the available data in this example, each of these latent variables would be matched with a single indicator, being the observed manifestations of those properties. In this paper we will demonstrate the principles of composites while using the simpler case where the indicators $x_1 - x_5$ are assumed to be measured without error. It should be understood by the reader that the method of composites can be extended to the case where latent variables are used to represent the true values of all our measured properties.

Our discussion of the range of possible ways to represent the relations in Figure 5 would be incomplete without considering another alternative, which we will refer to as Model C (fig. 8). In this model, our two multidimensional constructs, soil conditions and landscape properties, are represented simply as groups of variables.

Formal Representations

To formalize our discussion of the relationships embodied in our models, we present the three characteristic equations of the LISREL model (Bollen, 1989);

$$x = \Lambda_x \xi + \delta, \quad (1)$$

$$y = \Lambda_y \eta + \varepsilon, \quad (2)$$

and

$$\eta = B\eta + \Gamma\xi + \zeta, \quad (3)$$

where x and y are vectors of observed indicators of exogenous (predictor) and endogenous (response)

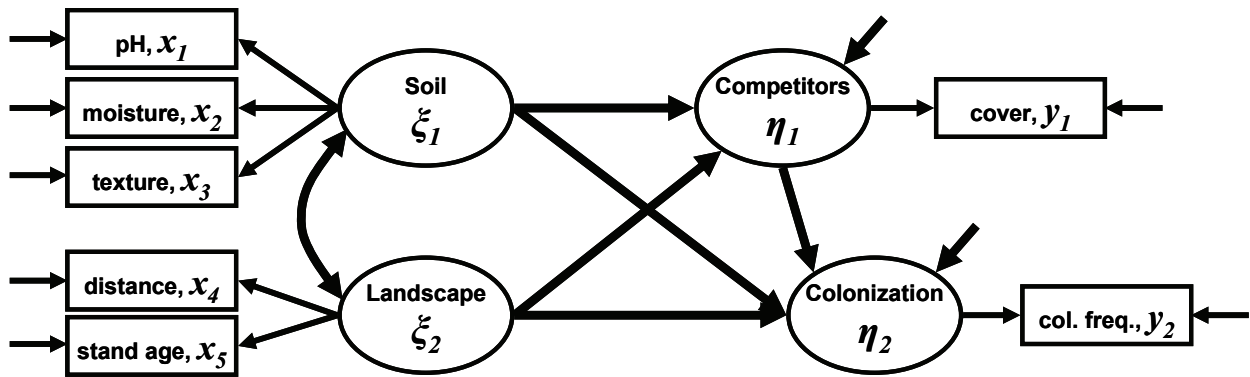


Figure 6. Model A. All observed (manifest) variables (in boxes) are represented as “effect indicators” associated with latent variables (in ovals). Thus, all blocks are of the $L \rightarrow M$ type. “col. freq.” refers to colonization frequency.

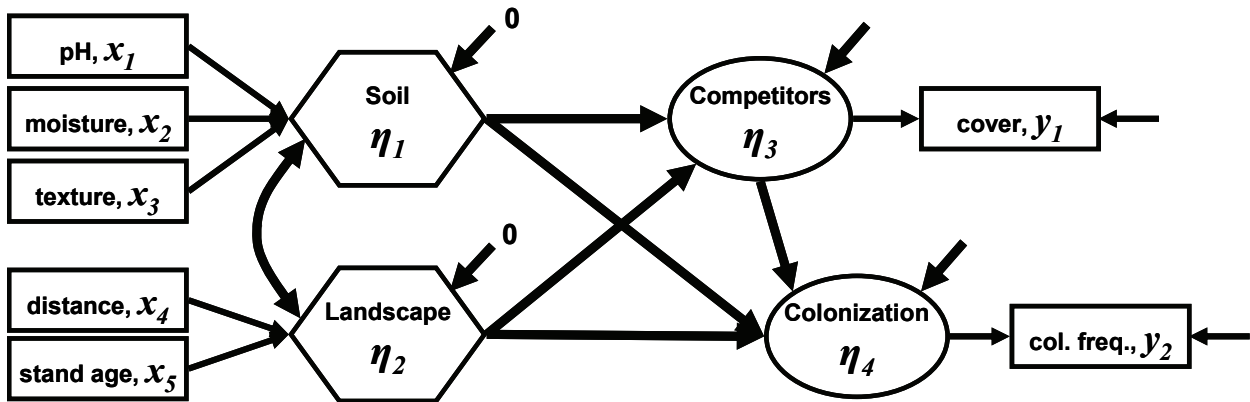


Figure 7. Model B. Soil conditions and landscape properties are represented as components of $M \rightarrow L$ blocks, with the associated manifest variables ($x_1 - x_5$) representing causal indicators. “col. freq.” refers to colonization frequency.

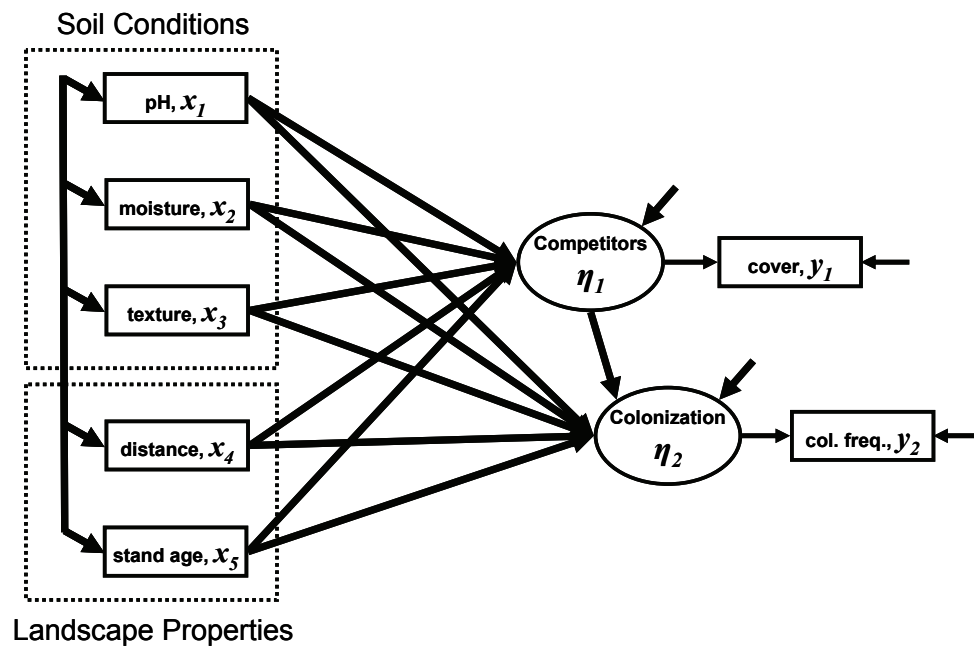


Figure 8. Model C. A partially reduced form of Model B omitting explicit consideration of collective soil and landscape effects.

latent variables, ξ and η are vectors containing the individual exogenous and endogenous latent variables, A_x and A_y are vectors of coefficients relating indicators to latent variables, B and Γ are coefficient matrices for effects of endogenous and exogenous latent variables on endogenous latent variables, δ and ε are vectors of measurement errors for x and y , and ζ is a vector of errors for the η variables. For exogenous latent variables, ξ , their variances are represented by the diagonal elements of the matrix Φ while the off-diagonal elements of the matrix are the covariances. The variances of the error terms for the endogenous latent variables (ζ) are contained within the diagonal elements of the Ψ matrix, while the off-diagonal elements of that matrix represent any covariances among errors (normally assumed to be zero). Error variables (in δ and ε) are expected to be uncorrelated with ξ and η . However, it is possible to model correlations among errors, typically represented by nonzero off-diagonal elements in the matrices representing cross products among measurement errors, θ_δ and θ_ε .

This notation is sufficiently general that we can use it to express the various models with causal indicators and composites that we discussed previously. We can incorporate the causal indicators in a model such as Model B by defining a separate ξ variable for each causal indicator and setting its factor loading to 1.0 and its error to zero. For instance, if x_1 to x_3 are causal indicators, we would have

$$\begin{aligned} x_1 &= \xi_1 \\ x_2 &= \xi_2 \\ x_3 &= \xi_3 \end{aligned} \quad (4)$$

as the part of the measurement model that corresponds to the causal indicators. This representation explicitly shows that we are assuming no measurement error in the x s.

Composites are included by specifying the coefficients in Γ that correspond to the variables that make-up the composite. So if the causal indicators from equation 4 form a composite and η_1 is the composite, we would estimate or fix the values in the first three columns of the first row of Γ to create the η_1 composite.

Estimation of Composites

Previous authors have typically estimated composites with nonzero error variances (e.g.,

MacCallum and Browne, 1993). In a sense, such composites are actually types of latent variables in that they are not fully measured by the data. As a practical matter, for such “latent composites” in blocks, such as $M \rightarrow L$ and $L \rightarrow L(\text{no } M)$, we usually have insufficient information to estimate the true variance for the latent variables since they do not have at least one effect indicator. In these situations, we cannot estimate the model as originally formulated and we must look for alternatives. One possibility is to replace $M \rightarrow L$ with $M \rightarrow C$ and $L \rightarrow L(\text{no } M)$ with $L \rightarrow C$, where C refers to a composite variable. In the case of $M \rightarrow C$, the block describes a composite of manifest variables, while in the case of $L \rightarrow C$, the block describes a composite of latent variables. Such composites are feasible elements of models, being both estimable and straightforward to interpret. This contrasts with the problematic characteristics of composites having non-zero errors (MacCallum and Browne, 1980).

Composites are, in all cases that we describe, considered endogenous (η) variables but they are endogenous variables completely determined by the variables that influence them. If a composite does not fully represent the construct with which it is associated, we must recognize that it is an imperfect measure of that construct. With regard to the example of forest herb colonization, if stand age and distance from source constitute the predominant landscape features of importance for the model, then a composite formed from their effects will have a very general meaning. If stand age and distance from source are just two of many unique landscape properties of importance to herb colonization, our composite will simply represent their joint effects and will be a poor representation of a general landscape effect. Typically, unless an effort has been made to sample the many facets of a construct, any composite derived from the measures of that construct should be understood to represent the collective effects of its components, regardless of the label placed on that construct. In such a case, an error estimate of zero is appropriate. Despite their limitations, composites can be useful as a means of representing certain types of constructs, facilitating general conclusions, and managing model complexity.

Evaluating the Different Model Possibilities - Theoretical Considerations

Both theoretical information and empirical information contribute to an evaluation of the suitability of model architecture for a given situation.

As stated earlier, there are a great number of possibilities for how a set of manifest variables may be connected. In the example under consideration here involving herb colonization of forest stands, initial guidance comes from the theory specified in figure 4 and the associations in figure 5. We begin our evaluation of the applicability of Models A-C in this case by considering the causal relations between latent variables and indicators.

Borrowing, with modification, from Jarvis and others (2003), we recognize certain conceptual criteria for deciding whether a block should be modeled in the traditional $L \rightarrow M$ format (e.g., Model A) or whether $M \rightarrow C$ blocks are more appropriate (e.g., Model B). Three kinds of questions can be considered to gauge the theoretical basis for forming models such as Model A versus Model B: (1) What is the direction of causality? (2) Are the indicators in a single block interchangeable? (3) Is there an expectation that indicators in a block should covary (e.g., because they are jointly controlled by an underlying factor)?

For the first question, we must begin by asking whether Soil has causal influences on pH, moisture, and texture. In other words, we are asking whether the variation among stands in soil conditions is such that pH, moisture, and texture are controlled by a common difference among soils. In such a case, we view that causation flows from the general construct to the indicators. Alternatively, if it is the case that pH, moisture, and texture behave independently from one another, then causation flows from the indicators to the latent variable. In this study, Verheyen and others (2003) felt that causation flows from the indicators and could be represented by a $M \rightarrow C$ block. A similar determination was made for the latent variable Landscape which is envisioned by the authors to be the result of largely distinct influences by certain landscape properties.

The second question to be considered is whether the indicators in a block are interchangeable; if so, then they constitute redundant measures and are likely to represent effect indicators consistent with a $L \rightarrow M$ block. Stated in another way, we might ask if dropping one of the indicators in a block alters the meaning of the construct. For both Soil and Landscape, the indicators in the separate blocks would seem to be unique based on the Verheyen and others' (2003) model specification. For the construct Landscape, an $M \rightarrow C$ block specification would appear to be indicated based on this criterion. The age of a patch is conceptually quite distinct from its distance from a source population. If data for the distance to source populations was absent, the construct would reduce to solely an age dimension. For Soil, there are precedents for specifying common

soil factors of the $L \rightarrow M$ type (Grace, 2003; Weiher and others, 2004). In this case, Verheyen and others (2003) felt that the individual indicators (pH, texture, and moisture) reflect separate dimensions of Soil and are best represented by an $M \rightarrow C$ block structure.

The third question is whether the indicators are expected to covary. Addressing this question serves as an additional way of evaluating the flow of causation in a block. If indicators are under common causal control by a latent factor, then when that latent factor varies, the indicators should all likewise vary. Such a situation would imply that an $L \rightarrow M$ block would be an appropriate means of modeling the situation. On the other hand, if causation flows from the indicators to the latent variable, there is no basis for expecting correlations among indicators since their causes are not specified. We should, therefore, recognize that a correlation among indicators does not inform us as to the direction of causal flow, though a lack of correlation among indicators would contraindicate the prospect that a block should be of the $L \rightarrow M$ form. For the herb colonization example, Verheyen and others' (2003) presentation did not produce any expectations of whether indicators in multi-indicator blocks would correlate or not. So, again, by this criteria, it would seem that Soil and Landscape are constructs best represented as $M \rightarrow C$ blocks given the available indicators.

Collectively, it would seem that the expectations of Verheyen and others (2003) as reflected by the three questions would lean towards a specification of the multi-indicator blocks as $M \rightarrow C$ for both Soil and Landscape. This assessment does not guarantee that this is the correct causal structure nor does it override the possibility that empirical characteristics of the data might imply a different block structure. What is most important to realize is that there are logical tests for developing, justifying, and interpreting block structure and that automatically presuming a conventional model structure ($L \rightarrow M$) is not justified. Neither is it advisable to simply rely on data properties to make the determination as to whether a construct is best represented by a particular structure.

Evaluating the Different Model Possibilities - Empirical Evaluations of Blocks

It is often not known with absolute confidence what the causal flow is for a set of manifest variables. This condition is especially common for applications in the natural sciences where latent variables are only beginning to receive attention. In SEM practice, logical and statistical approaches to evaluating the appropriateness of model structure exist (see

numerous illustrations in Bollen, 1989). We will address overall model fit later. For now, our considerations relate to the empirical characteristics of individual blocks. Ultimately these are of great importance, for no matter what logic one brings to a model, if the indicators in a block do not behave like effect indicators, then it will be unprofitable to model them as such. Reviews of SEM studies in certain disciplines (e.g., Jarvis and others, 2003) suggest that researchers commonly misjudge the empirical support for the causal direction specified in their models. The greatest culprit is the assumption that blocks are of the $L \rightarrow M$ form without seriously considering the use of $M \rightarrow L$ and $L \rightarrow L$ (no M) blocks (or $M \rightarrow C$ and $L \rightarrow C$). In the example being discussed here, Verheyen and others (2003) assumed the form ($M \rightarrow C$) for their multi-indicator blocks. It seems, however, that the logical justification for this assumption is stronger for the Landscape construct and less definitive for the Soil construct. In this section, we consider some of the empirical properties that inform us as to whether either block has properties consistent with $L \rightarrow M$ form.

Table 1 presents correlations for colonization frequency and other manifest variables (note that the covariances, the basis for SEM analysis, can be reconstructed from the correlations and standard deviations) for the species *Lamium galeobdolon*. As Bollen and Ting (2000) have described, there is an expectation that a set of effect indicators associated with a single latent variable in a $L \rightarrow M$ block will be correlated, as implied by equation 1. For such a set of indicators, their degree of correlation will depend on the strength of the common influence of the latent cause relative to their total variances. So, for Model A, we would expect conspicuous correlations among soil pH, moisture, and texture because of the joint influence of Soil on those specific factors. We would have similar expectations for a strong correlation between distance and patch age. Again, the degree of correlation expected would depend on the relative importance of the individual errors for each indicator, though for a reliable set of indicators, correlations should be at least moderately high. In contrast, for a set of causal indicators associated with a single latent variable in a single $M \rightarrow C$ block, there is no basis for expecting any particular correlation pattern. None of the equations that apply to Model B imply common causal influence on sets of causal indicators. So, a set of causal indicators either may or may not intercorrelate in such a case, since our equations do not describe their interrelationships, except that they are classified as being of common interest.

Inspection of table 1 reveals a correlation of moderate magnitude between age and distance of

-0.5934. In addition, the correlations between these variables and the other variables in the model are approximately of the same magnitude. Thus, based on a crude inspection of correlations in the matrix, we are unable to rule out the possibility that either a $L \rightarrow M$ or $M \rightarrow C$ block structure would be consistent with the empirical characteristics of the indicators related to Landscape.

The correlations between pH, moisture, and texture are 0.0265, 0.1324, and 0.5767 respectively. The low magnitude of correlations between pH and the other indicators in the block suggests that these three soil properties would not be likely to represent redundant measures of the sort normally found in $L \rightarrow M$ blocks. A method for formally evaluating causal versus effect indicators has been proposed by Bollen and Ting (2000) based on a simultaneous analysis of vanishing tetrads in models.⁴

Correlations/covariances among a set of truly redundant indicators in a $L \rightarrow M$ block should possess the mathematical property of vanishing tetrads, with a tetrad being the difference between the products of pairs of covariances among four random variables. It can be said in this case that the pattern of correlations among pH, moisture, and texture does not appear to be consistent with such a block structure. To increase our confidence in this assessment, we applied Bollen and Ting's (2000) vanishing tetrad test to more formally compare Model A to Model B. Models B and C do not imply any vanishing tetrads whereas Model A does. We used the SAS macro for vanishing tetrads provided in Hipp and others (2005) and found that the fit of Model A is significantly worse than is the fit of Model B or C ($p < 0.05$).

Comparing Results for Different Models

Results for Model A

The data presented in table 1 are sufficient to provide for an evaluation of all models discussed in this paper. Analyses were conducted by using the program Mplus (Muthén and Muthén, 2004). We used maximum likelihood estimation in all analyses. For the purposes of this illustration, we assume a modest, nonzero degree of measurement error for the

⁴ Spirtes and others, 1993 have done important work in exploratory tetrad analysis, providing methods that can help to generate plausible model structures for a data set. Here we are interested in confirmatory tetrad analysis where we are testing specific hypotheses by using a simultaneous test statistic developed in Bollen (1990).

single-indicator measures (cover and colonization frequency) of 10 %.

The specification of a nonzero percentage of measurement error deserves discussion. Here we specify 10 % measurement error primarily to illustrate a common use for single-indicator latent variables. It is widely recognized among statisticians that the assumption of perfect measurement (zero measurement error) is frequently unjustified and a common source of bias for model parameters. In SEM practice, there exist both a readily available means of adjusting for known measurement error (the specification of measurement error estimates using latent variables) and a tendency to make such adjustments. When single-indicator latent variables are specified, one must specify the most reasonable estimate for the degree of measurement error. Often, though not always, this is some nonzero percent of observed variance. It is best if this estimate comes from empirical evidence, such as from replicate samples or a calibration data set. However, sometimes an estimate is made based on subjective reasoning, which, although approximate, is viewed as less arbitrary than assuming zero error and likely to produce a lower degree of bias in parameters than achieved from a zero error estimate. We should say that the specification of measurement error in this example is not a requirement for the use of composites, nor does it play a major role in our presentation. That said, those who choose to specify nonzero measurement error in SE models should be aware of the impacts on model results and their stability (Bollen 1989).

The fitting of data to Model A (fig. 6) resulted in poor fit based on model chi-square statistics, error variances for indicators, and examination of the residual covariance matrix. A chi-square of 45.20 was obtained for the model, with 10 degrees of freedom. The associated probability of good fit between data and model was found to be less than 0.00005. Note that since Model A is saturated with regard to the structural model (i.e., all paths between latent variables are estimated), the inflated chi-square does not result from unspecified relations among latent variables but instead resides in the measurement model. In other words, in this example the lack of model fit can be attributed to inappropriate relations between latent and manifest variables. Standardized parameter estimates (for alternatives to the use of conventionally standardized parameters, see Grace and Bollen, 2005) for Model A are shown in figure 9. Here it can be seen that the standardized error variance for pH is 0.99, indicating that the model provides no explanation of the variance of this variable. Finally, residuals indicate

that in general, the model does a poor job of resolving the covariances among indicators.

A modification of Model A was evaluated to determine the degree to which the lack of fit for pH contributed to poor model fit. In this modified model, only moisture and texture were used as indicators for Soil, while pH was specified to be a single indicator for an additional exogenous latent variable, True pH. Again, 10 % of the variance of observed pH was specified as being measurement error. A chi-square of 34.57 with 7 degrees of freedom was obtained for this model. The associated probability of good fit remained less than 0.00005. Thus, this result indicates that the observed lack of fit between model and data is spread throughout the measurement model and not solely due to a lack of correlation between pH and the other soil variables.

A Digression on Parameter Identification in Models with Composites

It is essential that we consider certain complications associated with the estimation of models containing composites if we are to overcome some of the limitations inherent in earlier presentations. These complications have specifically to do with being able to avoid problems of model underidentification. Model B (fig. 7), because it includes composite variables, has a somewhat different set of parameters that must be estimated compared to a similar model without composites (e.g., Model C, fig. 8). As mentioned earlier, since composites are defined by having zero error variances, the identification of the error variances for composites is not an issue. For Model B, associated with the composite variables are four paths that represent the effects of the composites on the latent variables. In this particular case, the relationships between the exogenous manifest variables (x s) and the composites, \bullet_1 and \bullet_2 , are reduced to single effects to each composite from each x . So, for our example, the 10 potential paths from the individual x variables to Competitors and Colonization in Model C are replaced in Model B with 5 paths from the x variables to the two composites plus 4 paths from the composites to Competitors and Colonization. Despite the net gain in degrees of freedom from this substitution, problems remain with parameter identification. Ultimately, there is a general problem that arises when attempting to identify all paths leading to, as well as flowing out from, a composite. This problem is similar to the routinely encountered problem associated with latent variables with effect

Table 1. Correlations among variables and the standard deviations. "col. freq" refers to colonization frequency.

	col. freq.	distance	age	texture	moisture	pH	cover
col. freq.	1.0						
distance	-0.5785	1.0					
age	0.6424	-0.5934	1.0				
texture	-0.2553	0.1844	-0.3146	1.0			
moisture	-0.3369	0.4604	-0.3462	0.5767	1.0		
pH	-0.0073	-0.0465	-0.0976	0.1324	0.0265	1.0	
cover	-0.3423	0.2070	-0.4062	0.3189	0.2952	0.2394	1.0
std. dev.	0.1867	0.8207	0.3390	0.4588	1.1359	0.7507	0.3017

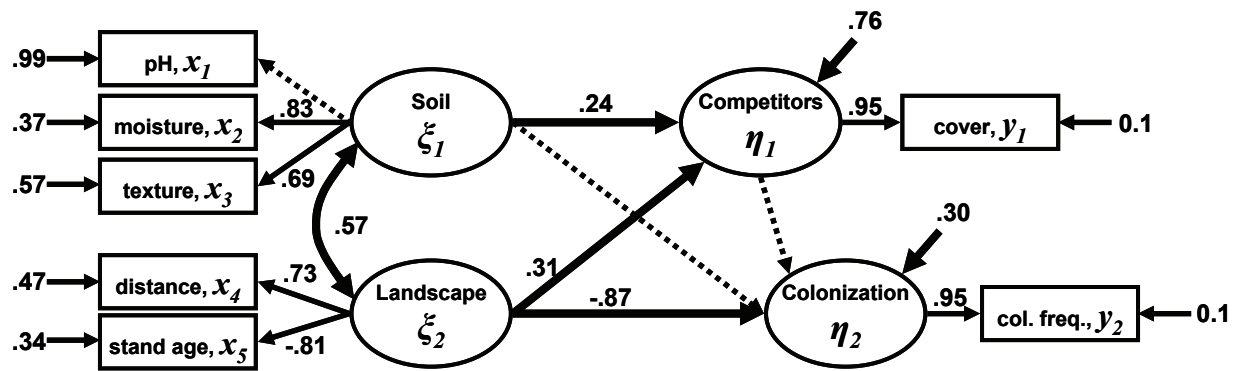


Figure 9. Standardized parameter estimates for Model A. Chi-square for model fit was 45.20 with 10 degrees of freedom (sample size = 180) and a p-value < 0.00005, indicating very poor fit of data to model expectations.

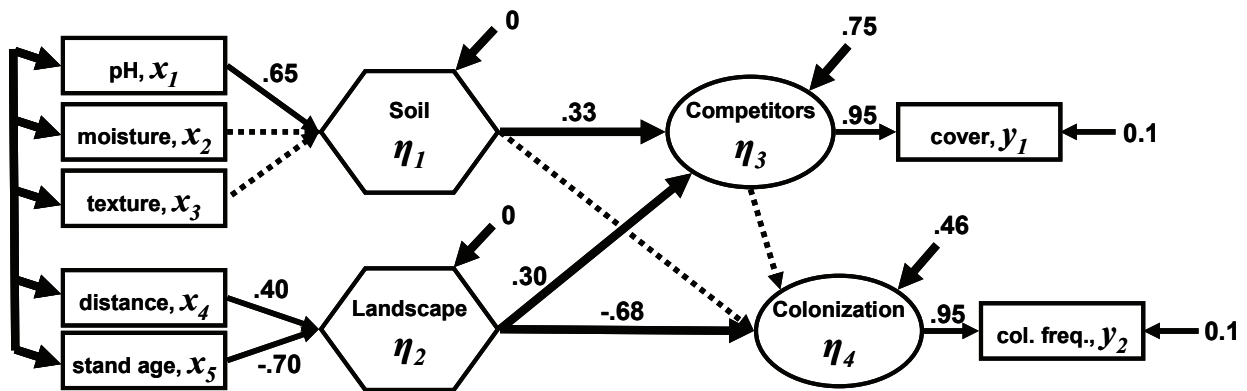


Figure 10. Standardized parameter estimates for Model B. Model chi-square was 6.888 with 3 degrees of freedom ($p = 0.075$).

indicators where the scale of the latent variable needs to be established. Both of these problems can be solved in the same fashion, by specifying a single incoming or outgoing relationship so as to establish the scale of measurement.

Despite the fact that parameter identification issues can be resolved by specification of select parameters, an issue still remains as to the statistical significance of the specified pathways. In Model B (fig. 7), we set the unstandardized coefficients for the paths from pH to Soil and from distance to Landscape to 1.0 to establish the scale of measurement for the composites (note that this is not shown in later results because only the standardized parameters are presented). This procedure ignores the question of whether the parameter values for these paths are detectably different from zero; specifically, does pH have a significant (nonzero) contribution to Soil, and similarly, does distance have a nonzero contribution to Landscape? One approach to evaluating fixed paths from x s to composites is to use a reduced form model such as Model C (fig. 8), in which the composites are omitted and the direct effects of x s on Competitors and Colonization are tested. For our example, the reduced form equations that apply to Figure 8 are

$$\eta_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \gamma_{13}x_3 + \gamma_{14}x_4 + \gamma_{15}x_5 + \zeta_1 \quad (5)$$

and

$$\eta_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \gamma_{23}x_3 + \gamma_{24}x_4 + \gamma_{25}x_5 + \beta_{21}\eta_1 + \zeta_2 \quad (6).$$

For models with more than a single path flowing out from the composites, as is the case in Model B, an evaluation of Model C only provides an approximate answer to the question of whether the parameters in Model B are significant. The reason the reason is only approximate is because the gammas relating x s to Competitors must be proportional to those relating the same x s to Colonization. In other words, it is assumed that the weightings specifying the contributions of pH, moisture, and texture to a composite of their effects on Competitors are the same as the weightings for the effects of those same variables on Colonization. This means that for the reduced form equations (5 and 6), $\bullet_{11} - \bullet_{15}$ should be proportional to $\bullet_{21} - \bullet_{25}$. Otherwise, a single coefficient will not, for example, adequately represent the paths from \bullet_1 to \bullet_3 and \bullet_4 , in Model B (fig. 7). As we show below, it is important to evaluate such relations as part of the assessment process.

Further Results

Results for Model B

Maximum likelihood estimation yielded a chi-square of 6.888 with 3 degrees of freedom and an associated probability of 0.075. Examination of residuals and the chi-square results indicate adequate fit of the data to Model B, though there were some slight indications of residual relations between distance and stand age and \bullet_3 and \bullet_4 . In the presentation of results in figure 10, all numerical estimates shown are associated with significant p values for the relevant unstandardized parameters, except for the paths from pH to Soil and distance to Landscape, which were set to fixed values to establish the scales for the composites.

Conspicuous in the results for Model B (fig. 10) is that neither moisture nor texture contributed significantly to the composite variable Soil. Also, Soil had a significant effect on Competitors but not on Colonization. In contrast, the relationship between stand age and Landscape was significant, as were both paths from Landscape to Competitors and Colonization. Examination of the residual variances for Competitors and Colonization shows that 25 % and 54 % of the variance in these response variables were explained ($R^2 = 1 - \text{the standardized error variance}$).

Results for Model C

It is instructive to evaluate Model C, which effectively represents what Bollen and Davis (1994) referred to as a “partially reduced form” of Model B in which composites are omitted. In Model C, separate effects of the x variables on Competitors and Colonization are estimated, and the individual x variables are associated with constructs nominally. Since this model is saturated, maximum likelihood estimation yields a chi-square of 0 due to the fact that there are no degrees of freedom. In a side analysis, the stability of results was tested by deleting nonsignificant paths, yielding a chi-square of 9.773 with 6 degrees of freedom ($p = 0.1344$). This side analysis indicated that the results for Model C are stable.

Results for Model C (fig. 11) show that only four of the paths from x variables to Competitors and Colonization were significant. Variances explained for Competitors and Colonization in the saturated

model were 30 % and 57 % respectively. Again, for the purpose of comparing among models, the saturated model results will be used even though nonsignificant paths are included. Aside from providing us with a comparison to Model B, Model C also provides a representation of our constructs from figure 4, except that in this case, the dimensions of soil conditions and landscape properties are represented separately.

A Further Model to Consider – Model D

Above it was mentioned that Model B presumes a single set of coefficients defines composite effects on more than one response variable. This is akin to saying that the slopes of the regression relations of x_s on y_3 would be proportional to the effects of x_s on y_4 . If this condition holds, then Model B is our best model for the situation. On the other hand, if this condition does not hold because relations are not proportional, our model will underestimate the variance explanation for Competitors and Colonization. It can be easily seen that, when gammas in equations 5 and 6 are not proportional, the composites in Model B will be insufficiently relevant to accurately predict both Competitors and Colonization accurately. To evaluate this possibility, we consider one additional model, Model D (fig. 12). In this model, separate composites are derived to estimate effects on Competitors and Colonization. Results for Model D were as for Model C in that when all paths are included there are no degrees of freedom for testing model fit. However, deletion of nonsignificant paths allowed for model testing, and again, no unspecified paths were indicated.

Comparisons Among Models

The exercise of comparing different models as contrasting representations of the construct relations in Figure 4 is meant not only to illustrate some of the variety of modeling possibilities but also to introduce a number of important concepts and procedures. Most of the presentations of SEM in introductory textbooks would likely lead beginning users to consider Model A to represent the construct relations in figure 4. In fact, several review papers have suggested that researchers frequently fail to consider the full range of possibilities for model development (Jarvis and others, 2003; Williams and others, 2003). Models that contain causal indicators and composites are rather rare compared to those in which constructs are exclusively represented by effect indicators in $L \rightarrow M$ blocks. In Example 1, the bulk of the

theoretical information and the great majority of the empirical analyses indicate that Model A is not an appropriate representation of the situation.

Conceptually, the constructs were not envisioned as unidimensional underlying causes but rather, as collections of influencing factors. Further, the available indicators do not represent a suite of effect indicators as would be appropriate for an $L \rightarrow M$ block. The inadequacy of fit of data to Model A is clearly indicated by the very low probability associated with the chi-square test. Thus, we can reject this model as inadequate. Considerations of modifications such as splitting out pH from the Soil block, as well as other attempts at small modifications, did not lead to the discovery of an adequate model akin to Model A either.

The remaining comparisons of most interest are those between Models B, C, and D. All three of these models had adequate fit to the data, and the differences among them can be viewed as tradeoffs between model generality and model accuracy (Levins 1968). Model B is the most conceptually general of the three (though it is more statistically restrictive), representing the collective effects of soil conditions and landscape properties with a single composite for each. As described in the above discussion, for Model B to be appropriate, the composites estimated would need to be near optimal for representing effects on both Competitors and Colonization simultaneously. At the other end of the spectrum, Model C is the most accurate and least conceptually general of the three models.

Compared to Model C, which has no composites, Model D represents an advance in generality, having an individual composite for each effect of the multidimensional constructs. As would be expected for such a case, the error variances for Competitors and Colonization are virtually identical for Models C and D. Thus, Models C and D are both equally accurate with regard to variance explanation, but they differ in the ways in which effects are represented. Overall, Model C provides a more succinct representation of the individual effects of x variables on Competitors and Colonization. All paths are tested for significance, and coefficients are readily interpretable as individual effects. In comparison, Model D provides a superior representation of collective effects, with single path coefficients representing the effects of suites of variables.

An important comparison to make between Models B and D has to do with the magnitudes of the error variances for the response variables. It has been recognized previously that, when composites with multiple effects on other variables are not

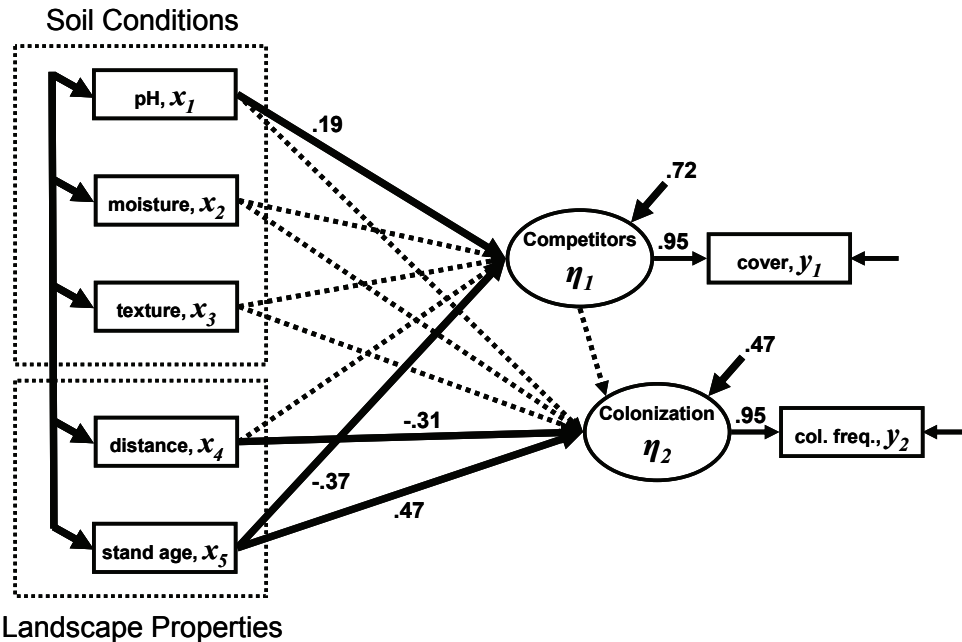


Figure 11. Select standardized parameter estimates for Model C. Here composites are omitted, thus, Model C represents a partially reduced form of Model B.

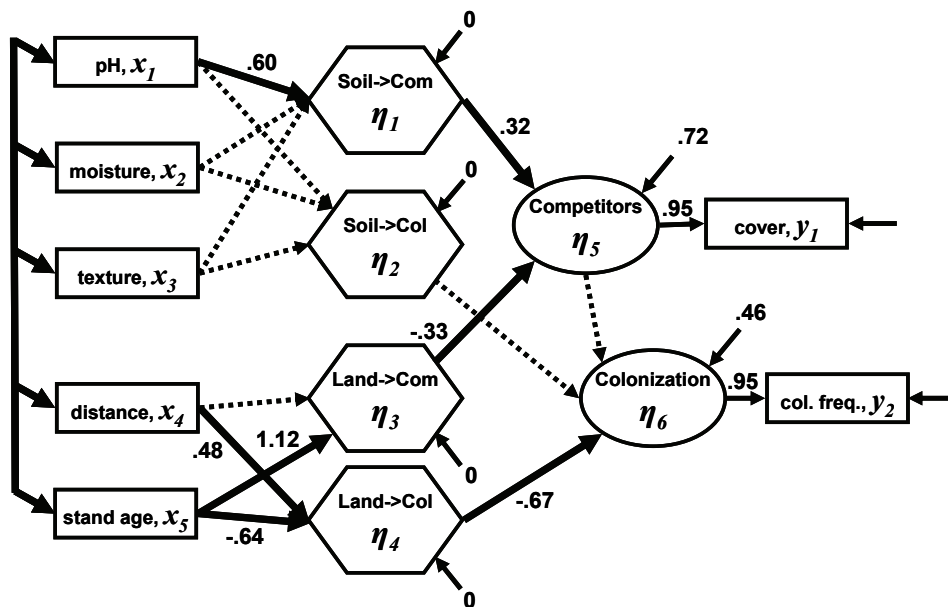


Figure 12. Standardized parameter estimates for Model D. Here four composites (recognized by the zero error variances) are included, one for each response variable.

appropriate, there is a conspicuous loss of variance explanation for the response variables. In this case, error variances are 0.75 and 0.46 for Competitors and Colonization, respectively, in Model B. In contrast, in Model D, the values are 0.72 and 0.46. A test of the adequacy of Model B is whether its error variances are significantly greater than those for the less general case where composites have single effects. For such a test, the t-value for a test of difference between the error variances for Competitors must be obtained. For Model B, the unstandardized residual variance for Competitors is 0.059 ± 0.007 (standard error), while for Model D the value is 0.057 ± 0.007 , yielding a t-value of 0.14 and an associated probability of no difference approaching 0.9. For Colonization, residual error variances were virtually identical (0.013 ± 0.002) for Models B and D. Thus, there is no indication that Model D is superior to Model B in variance explanation for the response variables. Further, the paths from composites to the response variables are also very similar between models. Collectively, this evidence supports the conclusion that Model B is superior to Model D overall because of its greater conceptual generality and that Model B is our best model (of those examined) for this example and these data.

Composites in More Complex Situations

In the example above, only the exogenous observed variables were related to composites. Situations can exist where endogenous constructs are best represented by composites. In this case, all the above material applies, but there are some additional complications to model structure that apply. An example of the evaluation of such a case is given in this section, dealing with the associations between anuran species richness and habitat conditions. In this example, it can be seen that the principles presented here can be readily extended to include a broader range of model types than the one considered above.

Example Involving Endogenous and Exogenous Composites: Examination of Habitat Associations for Anurans

Background

Anurans are a group of amphibians, comprising frogs and toads, that are of high conservation concern because of worldwide population declines. Example 2 uses data from a study by Lichtenberg and others (2006) in which 25 wetlands in the lower Mississippi River alluvial valley were examined for chorusing anuran species and associated habitat characteristics. One major goal of this study was to understand the habitat characteristics associated with diversity “hotspots,” places where a variety of species are abundant. Figure 13 provides an illustration of the construct relations of interest in Example 2. Stated in general terms, the goal of their analysis of anuran diversity was to understand the degree to which it depends on the type of habitat (lake, impoundment, etc.) versus the particular conditions (vegetation, litter, etc.) within that habitat. Specifically, the investigators were interested in the question of whether correlations between high diversity and particular wetland types can be explained by the microhabitat conditions within the wetlands. Such information could prove useful in making decisions about types of habitats to protect and the conditions to maintain.

In the study by Lichtenberg and others (2006), several types of wetlands were examined. These could be classified as being of one of four types – lakes, impoundments, swales, and riverine areas. Specific microhabitat conditions were assessed at each site, and these included a number of vegetation and topographic features. Anuran diversity was assessed by using nighttime surveys of chorusing individuals during several seasons of the year. Chorus surveys during the year were used to produce an estimate of the species using a site.

While macrohabitat types were clearly defined a priori, microhabitat conditions were sampled in a more exploratory fashion. A wide range of characteristics of the vegetation, including herbaceous cover, vegetation density at different vertical positions, woody cover, tree heights, canopy cover, as well as litter cover and depth (by type) were

measured. Also measured were hydrologic features at the site, including the area of open water and the mean and maximum water depths.

Example 1 provided a detailed consideration of how constructs may be represented and both theoretical and empirical criteria for arriving at decisions about block structure. In Example 2, our emphasis is more on the question of how to model a situation where an endogenous variable (microhabitat conditions) has multiple indicators and may involve composites. This question was not addressed in Example 1, where multiple indicators existed only for the exogenous constructs.

In this second example, we begin our analysis with the construct labeled Macrohabitat Type. Since our measure is nominal and multilevel (whether a site is classified as lake, impoundment, swale, or riverine), it immediately suggests the need to model this construct by using a set of dummy variables representing the possible macrohabitat types. We assume for the sake of simplicity that the classification of individual sites as to habitat type was correct. Therefore, our presumption is that the construct Macrohabitat Type can be modeled by using the $M \rightarrow C$ block type.

The construct labeled Microhabitat Conditions is one where the specific details of how the measured variables are related was not known *a priori*. For this reason, Lichtenberg and others (unpub. data) performed an exploratory factor analysis to see if the correlations among the many measured variables might suggest the operation of a smaller number of latent factors. Here we will not go into the details of that analysis but only say that the result was the recognition by the authors of two factors of importance to anuran diversity, the abundance of herbaceous vegetation and the abundance of leaf litter. Based on the conceptualization of Microhabitat Conditions by Lichtenberg and others (2006), it is clear that the indicators could represent a collection of factors that affect anurans based on theoretical grounds. Thus, we begin with the expectation of a block structure of $L \rightarrow C$, with two latent variables, Herbaceous Vegetation and Litter, contributing to the construct. A total of seven indicators of the two latent variables were included in their final model (see below).

Lichtenberg and others (2006) discuss certain issues of measurement regarding the construct Anuran Diversity. It is widely held that there are several causes of measurement error for wildlife populations and communities. In addition to the usual matter of sampling, varying detectability can contribute to error. Lichtenberg and others (2006) addressed the issue of detectability to some degree by

using the total number of species recorded across samplings instead of the mean. Nonetheless, error in assessing the true number of species at each site is likely significant and, while no estimate of this error exists, we again use an arbitrary estimate of 10 % of the total variance.

Possible Model Structures

In the first example we spent some time discussing and illustrating criteria for evaluating the cause and effect relations among variables and how these influence model structure. In this second example, we forego a detailed discussion of such considerations and, instead, focus on the possible ways to model a case where composites have directional effects on other composites.

A model that logically follows from the construct relations in figure 13 is shown in figure 14. We refer to this model as Model E to avoid confusion with the models discussed in conjunction with the first example. Based on the information presented above, we represent Macrohabitat by using the block structure $M \rightarrow C$. The Microhabitat construct is represented in this case as an $L \rightarrow C$ block containing two latent variables, both with multiple indicators. Diversity is represented by an $L \rightarrow M$ block with a single indicator. Consistent with their declarations as composites, the error variances for Macrohabitat and Microhabitat are set to zero.

Of the models we consider here, Model E represents the one that is most abstract. At the same time, Model E is based on the greatest number of assumptions. It is presumed in this model that the influences of Macrohabitat (\bullet_1) on Microhabitat (\bullet_2) can be summarized by a single coefficient, \bullet_{21} , despite the fact that \bullet_2 is actually a composite of latent effect of Herbaceous and Litter effects on Diversity rather than the microhabitat conditions themselves. Stated in other terms, in Model E the covariances between $x_1 - x_3$ and $y_1 - y_7$ must be resolved by their joint relations to a linear predictor that depends on \bullet_3 . These are fairly critical assumptions in that their failure likely means an underestimate of Macrohabitat effects on Microhabitat and possibly unresolved covariances among manifest variables.

An alternative formulation that is perhaps more biologically meaningful, Model F (fig. 15), represents the modeling implications of relaxing the assumptions just discussed for Model E. Here we are allowed to consider effects of Macrohabitat on the two dimensions of Microhabitat, Herbaceous and Litter. Since these two dimensions are independently

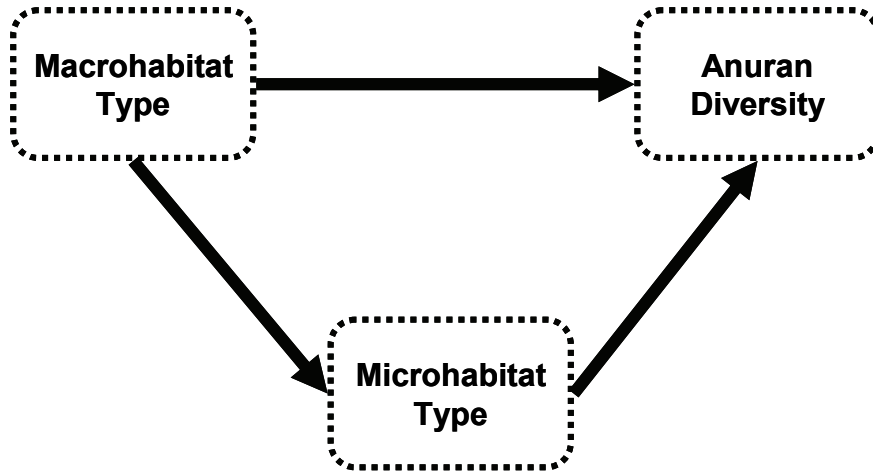


Figure 13. Construct model relating macrohabitat type and microhabitat conditions to anuran diversity.

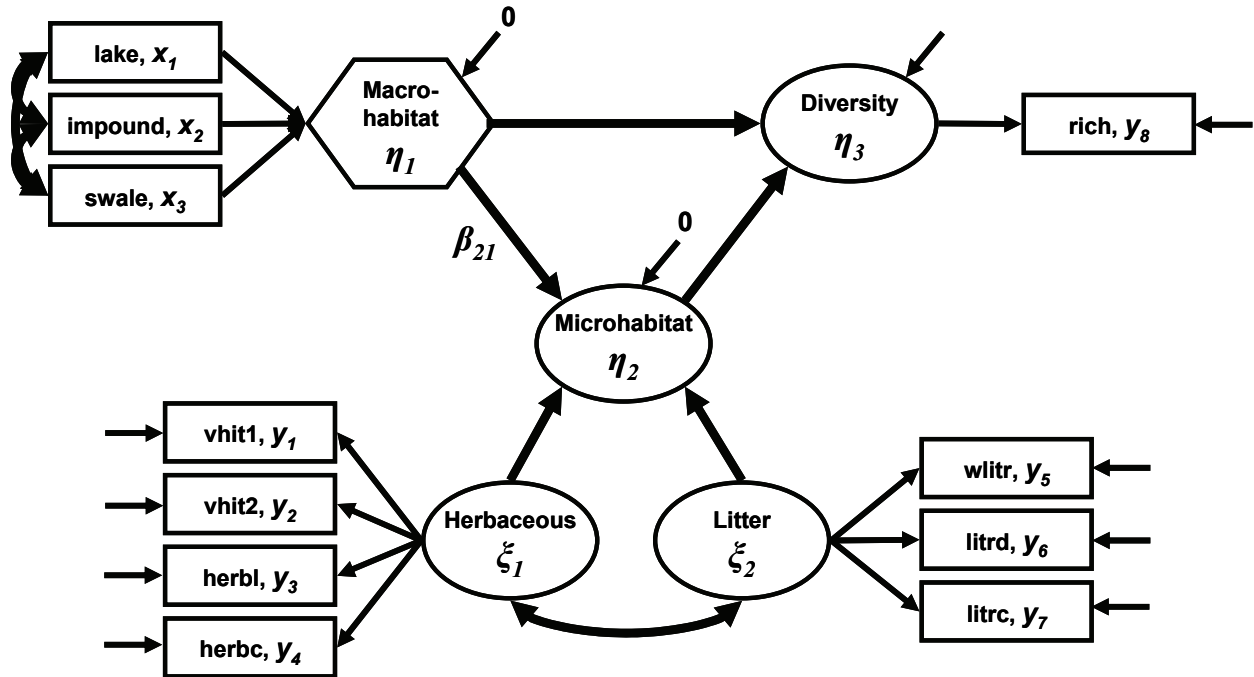


Figure 14. Model E, which shows one of the possible ways that indicators could be related to the construct model shown in Figure 13. Note that by omission, the riverine macrohabitat condition represents the baseline against which other macrohabitats are compared. “rich” refers to number of anurans at a site, “imp” refers to impoundments, “vhit2” and “vhit1” are measures of vegetation density, “herbl” and “herbc” are measures of dead and live herbaceous vegetation, “wlitr” is woody litter, “litrd” is litter depth, and “litc” is litter cover.

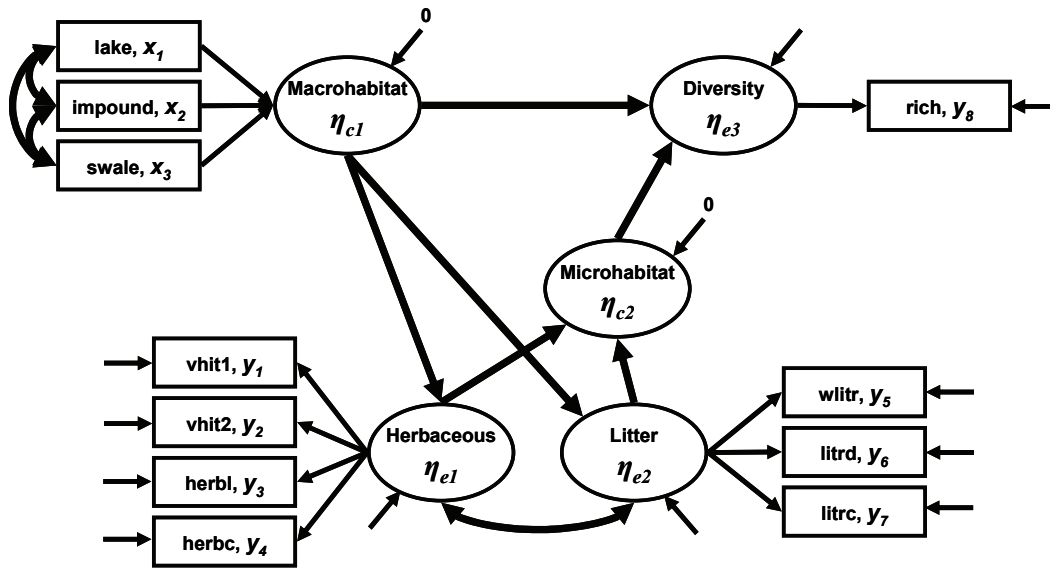


Figure 15. Model F, which differs from Model E by allowing separate coefficients (represented by separate paths) to convey the effects of Macrohabitat on the Herbaceous and Litter dimensions of Microhabitat (see caption for fig. 14 for abbreviations).

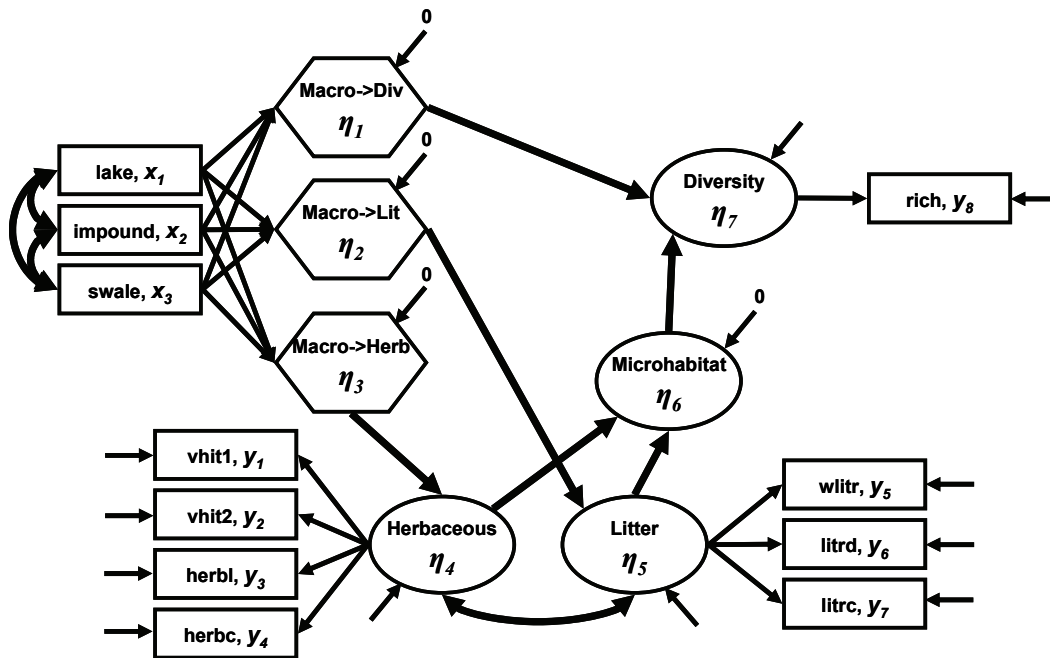


Figure 16. Model G, which differs from Model F by using separate composites ($\bullet_1 - \bullet_3$) to convey the effects of individual Macrohabitat types on \bullet_4 , \bullet_5 , and \bullet_7 (see caption for fig. 14 for abbreviations).

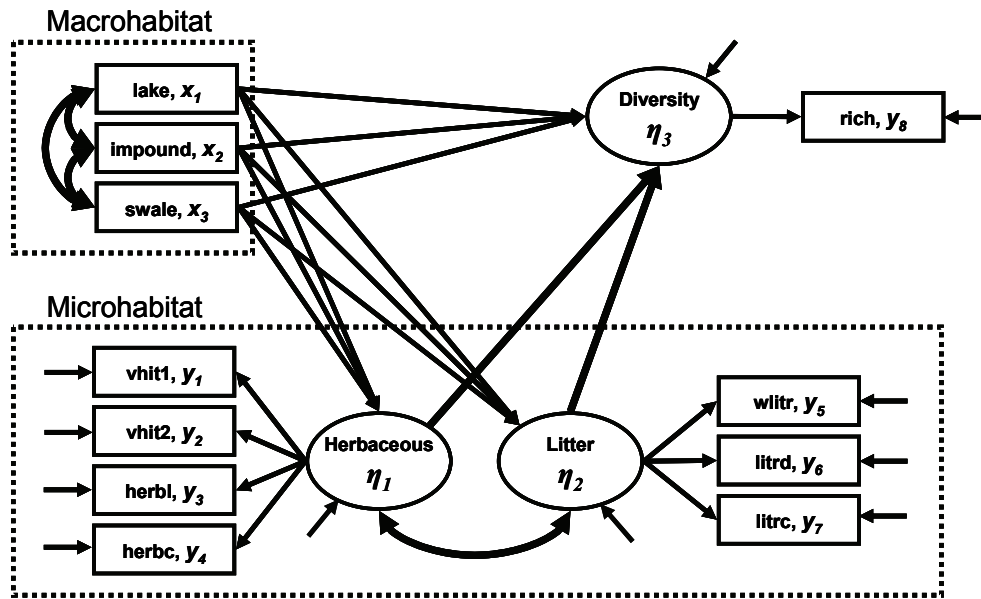


Figure 17. Model H, partially reduced form model representing effects of Macrohabitat and Microhabitat nominally through grouping variables (see caption for fig. 15 for abbreviations).

Table 2. Correlations among variables related to anuran richness and their standard deviations. “rich” refers to number of anurans at a site, “imp” refers to impoundments, “vhit2” and “vhit1” are measures of vegetation density, “herbl” and “herbc” are measures of dead and live herbaceous vegetation, “wlitr” is woody litter, “litrd” is litter depth, and “litc” is litter cover. “sd” refers to standard deviation.

	rich	lake	imp	swale	vhit2	vhit1	herbl	herbc	wlitr	litrd	litrc
rich	1.0										
lake	.696	1.0									
imp	-.167	-.355	1.0								
swale	-.431	-.659	-.253	1.0							
vhit2	.372	.167	.111	-.099	1.0						
vhit1	.222	-.156	.552	-.118	.653	1.0					
herbl	.060	-.252	.562	-.009	.581	.825	1.0				
herbc	.091	-.087	.419	-.132	.437	.745	.756	1.0			
wlitr	.509	.430	-.284	-.099	-.051	-.290	-.395	-.396	1.0		
litrd	.238	.146	-.433	.383	.027	-.097	-.180	-.281	.419	1.0	
litrc	.219	.194	-.442	.273	-.118	-.414	-.509	-.580	.568	.762	1.0
sd	2.170	0.510	0.332	0.476	0.512	1.482	0.173	0.122	0.100	0.122	0.148

estimated in L→M blocks, the interpretative meaning of the paths from Macrohabitat to each dimension is clear and also independent of other relationships in the model. In contrast, since the composite Microhabitat is defined as a linear predictor of Diversity, it depends on the covariances between three latent variables, \bullet_2 , \bullet_3 , and \bullet_5 . If any of the covariances among these variables changes, the meaning of \bullet_4 changes and, thus, the meaning of a direct path from Macrohabitat to Microhabitat (as in Model E). In Model F, the effect of Macrohabitat on Microhabitat can be summarized by calculating the total effect of \bullet_1 on \bullet_4 , which can be summarized by the equation

$$\beta_{41} = \beta_{21} * \beta_{42} + \beta_{31} * \beta_{43}, \quad (7)$$

where \bullet_{41} now refers to a calculated total effect.

Certain restrictive assumptions that warrant discussion remain in Model F. In particular, implied is the assumption that a single combination of coefficients from macrohabitat types ($x_1 - x_3$) to the Macrohabitat composite (\bullet_1) simultaneously summarizes the individual effects of $x_1 - x_3$ on \bullet_2 , \bullet_3 , and \bullet_5 . We feel that it is always wise to check such an assumption (as was done in Example 1). Here, this assumption is explicitly removed in Model G (fig. 16), where one composite representing Macrohabitat effects is replaced by three (Macro→Herb, Macro→Lit, Macro→Div).

Also as was done in Example 1, we offer for comparison a partially reduced form model, Model H (fig. 17), in which composites are omitted and constructs are represented nominally. While Model H offers a less complete modeling of the construct relations (e.g., there are no path coefficients derived for interactions among the major constructs), it permits a complete evaluation of the significance of all individual effects, thus complementing the other models.

Comparisons Among Models

Sample correlations and standard deviations are presented in table 2 for the manifest variables considered in this example. As stated earlier, we forego an empirical evaluation of the relations between manifest variables and either composites or latent variables. We also bypass any discussion of the bivariate relations in Table 2 and move directly into a consideration of model results.

As was brought out in our discussion of Example 1, the solution and evaluation of general models containing composite variables often depend

on information from more specific models. For this reason, we will begin with the model containing the least number of assumptions (and thereby, our least general model), Model H. We then will proceed progressively through models G, F, and E to determine whether more general models result in a substantial loss of information.

Results for Model H

The results for Model H are given in figure 18. For the purposes of this evaluation (and in contrast to our practice in Example 1), nonsignificant effects of macrohabitat types on endogenous variables were dropped from the final model. Despite the small sample size, results were stable, and the fit between model expectations and data was acceptable. As these results show, impoundments had significantly higher levels of herbaceous vegetation than did other macrohabitat types. Litter accumulation, in contrast, was substantially higher in lakes and swales than in impoundments and riverine habitats (recall that the riverine variable was omitted and, therefore, serves as the baseline condition). Anuran diversity was found to be higher in lakes than in all other habitat types.

Results for Model G

Model G provides for a single path from each composite to replace the multiple paths that would otherwise connect one construct with another. Aside from that, the models are very similar. Results from the estimation of Model G are presented in Figure 19. A comparison of results from Models H and G show numerous similarities, as well as a few differences. Model fit parameters are identical for the two models. Also, variance explanation for endogenous latent variables is the same, with R^2 s of 0.33, 0.39, and 0.78 for \bullet_4 , \bullet_5 , and \bullet_7 , respectively. Loadings in L→M blocks are the same for both models, as are outward paths from composites possessing single causes ($\bullet_1 \rightarrow \bullet_7$ and $\bullet_3 \rightarrow \bullet_4$) in comparison to the equivalent effects in Model H ($x_1 \rightarrow \bullet_3$ and $x_2 \rightarrow \bullet_1$). Composites with multiple causes in Model G yielded parameters not found in Model H, such as those associated with $\bullet_6 \rightarrow \bullet_7$ (Microhabitat to Diversity) and $\bullet_2 \rightarrow \bullet_5$ (Macro→Lit to Litter). The path coefficients associated with these paths represent standardized collective effects of the composites' causes on the response variables involved. Heise (1972) referred to these coefficients as "sheath" coefficients to designate the fact that they represent a collection of causes.

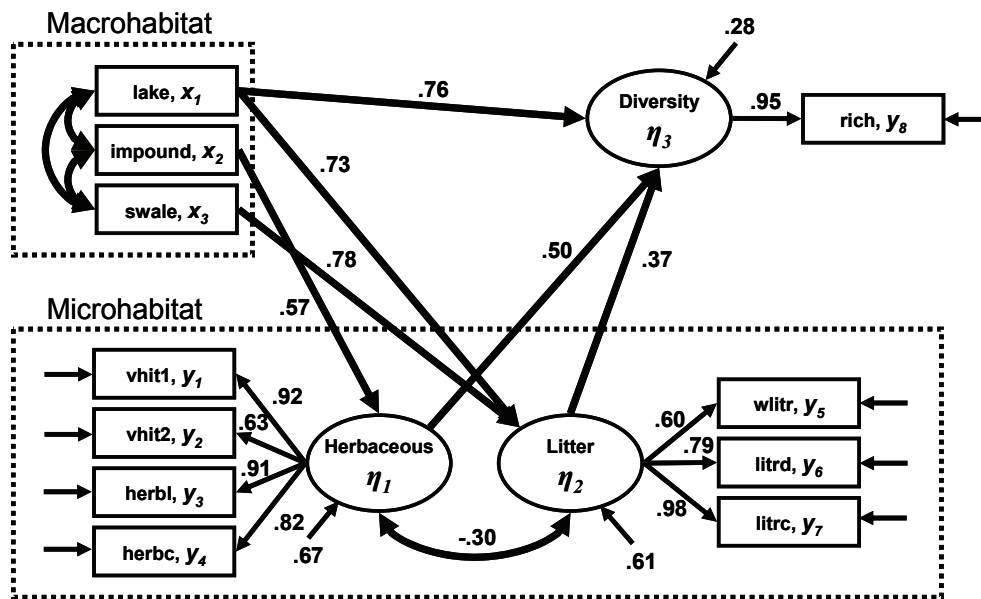


Figure 18. Results obtained for Model H, showing standardized values for path coefficients and the error variances of latent endogenous variables. Correlations among x s and errors for y s are not shown for simplicity. Nonsignificant effects of macrohabitat types were dropped from the final model, which possessed a chi-square of 50.42 with 39 degrees of freedom and a p of 0.104 (see caption for fig. 15 for abbreviations).

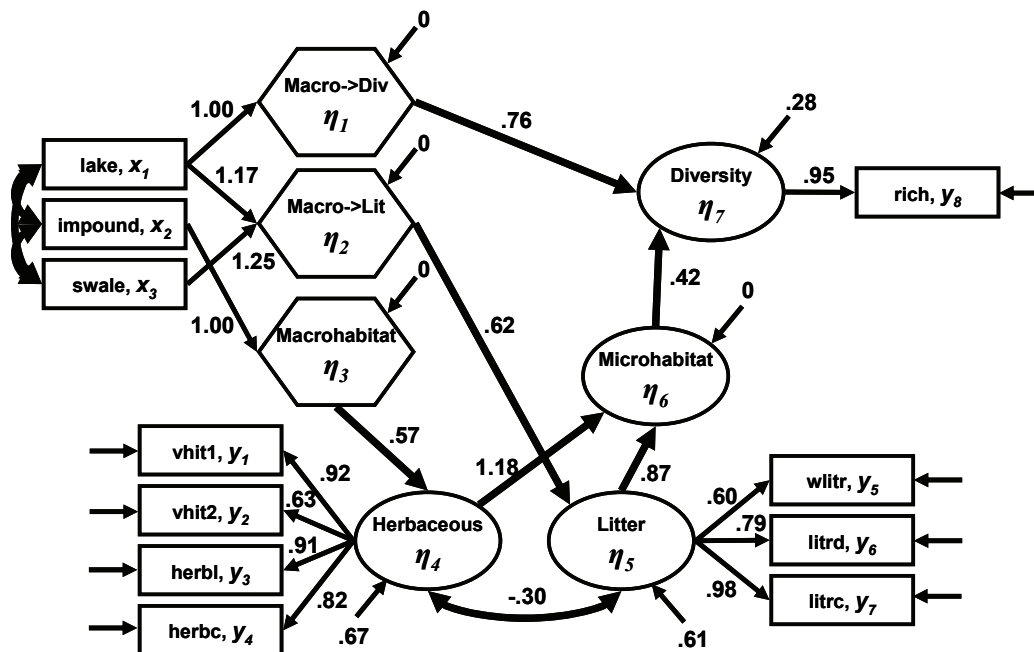


Figure 19. Results obtained for Model G, showing standardized values for path coefficients and the error variances of latent endogenous variables. Correlations among x s and errors for y s are not shown for simplicity. Nonsignificant effects of macrohabitat types were dropped from the final model, which possessed a chi-square of 50.42 with 39 degrees of freedom and a p of 0.104, precisely as for Model H (see caption for fig. 15 for abbreviations).

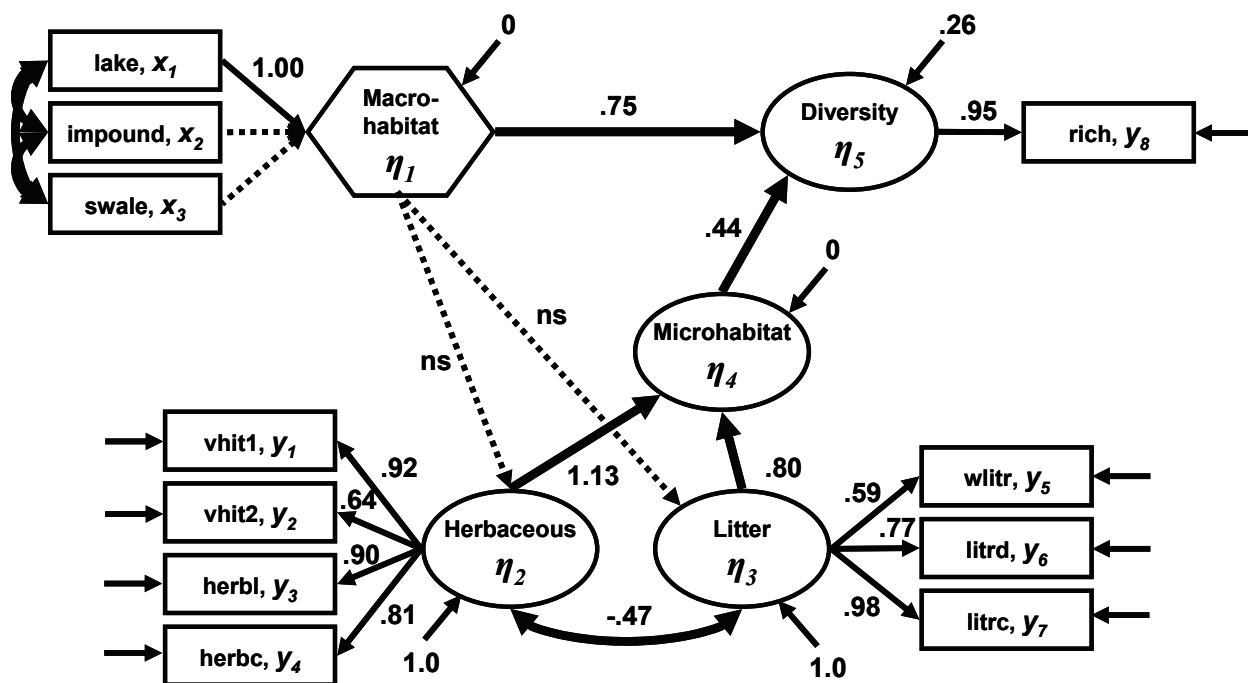


Figure 20. Results obtained for Model F, showing standardized values for path coefficients and the error variances of latent endogenous variables. Correlations among x s and errors for y s are not shown for simplicity. Chi-square for this model was 30.03 with 24 degrees of freedom and a p of 0.1837 (see caption for fig. 15 for abbreviations).

The most conspicuous differences between Models H and G reside with the paths from causes to composites. For example, in Model H, the effects of Herbaceous and Litter on Diversity are represented by two path coefficients (0.50 and 0.37), while in Model G, the same effects are represented by three paths, two from Herbaceous and Litter to Microhabitat (1.18 and 0.87) and one from Microhabitat to Diversity (0.42). Upon first examination, the paths from Herbaceous and Litter to Microhabitat appear unusually inflated, particularly since Herbaceous and Litter are only modestly correlated (-0.30). The equivalency of relationships in Models H and G can be made clear by realizing that the total effects of Herbaceous and Litter on Diversity are simply compound paths in Model G (e.g., the effect of \bullet_4 on $\bullet_7 = 1.18 \text{ times } 0.42 = 0.50$). The same holds true for all effects involving composites (e.g., the effects of x_1 and x_3 on \bullet_5).

Results for Model F

Model F, as described above, is a more general statement than Model G and presumes that a single composite is sufficient to represent the effects of Macrohabitat types on \bullet_2 , \bullet_3 , and \bullet_5 . Results from maximum likelihood estimation of Model F included a chi-square of model fit of 66.12 with 38 degrees of freedom and an associated p-value of 0.0031, indicating poor fit. It was found that when estimated by using a single composite for Macrohabitat effects, substantial residual correlations between individual macrohabitat types and the indicators for \bullet_2 , \bullet_3 , and \bullet_5 were not resolved. These results suggested that a single composite representing the effects of macrohabitat types on Herbaceous, Litter, and Diversity was inconsistent with the data. However, neither impoundment or swale indicators contributed significantly to the composite in this case; therefore, it was possible to simplify the model by dropping these indicators from the model. With these indicators deleted, model fit improved to 30.03 with 24 degrees of freedom and a p-value of 0.1837, indicating acceptable fit. The results presented in figure 20 are based on these results.

The version of Model F presented in figure 20, while representing a model that fits the data, fails to explain significant variation in either \bullet_2 or \bullet_3 . We might choose to select this model to represent the direct and indirect effects of Macrohabitat because it addresses the central question of interest in Lichtenberg and others' (2006) study. While it succeeds at that more narrow objective, Model F fails to provide an explanation for macrohabitat effects on both diversity and microhabitat. The clear reason for

that failure is that macrohabitat effects on microhabitat conditions are quite different from those on diversity, and that a single composite would not be able to represent both adequately.

A Consideration of Model E

Our most abstract model, Model E, represents direct effects of Macrohabitat on Microhabitat rather than on the dimensions of Microhabitat (Herbaceous and Litter). We found that a proper estimation of this model using a simultaneous solution procedure could not be achieved. The fundamental difficulty is that a zero error variance is specified for Microhabitat in order to permit it to represent the effects of Herbaceous and Litter on Diversity. It is not possible to estimate values for Microhabitat, a composite, based on the effects of Macrohabitat when the latter is also a composite. Thus, with Microhabitat represented as a composite with zero error variance, Model E is unresolvable without resorting to the derivation of a composite index for Microhabitat (such an approach would require a piecewise solution procedure, which we do not prefer). This condition serves as an additional argument in favor of modeling approaches in which exogenous influences are measured against the individual dimensions of endogenous multidimensional constructs (e.g., Models G and F) rather than the composite of their combined effects (in this case, on Diversity). With such an approach, effects on the endogenous composite can be estimated by using compound paths.

In conclusion, the principles presented by Grace and Bollen (in press) can be extended to models that contain endogenous composites. Some additional complexities arise with regard to model structure. In particular, the effects of composites representing exogenous indicators on composites representing endogenous indicators will typically need to be viewed as being indirect, through effects on the endogenous indicators themselves. This requirement poses no real problems for the inclusion of composites in such models, and our analysis is meant primarily to make clear the logic whereby effects in such models are considered.

Conclusions and Recommendations

Ecological theories often deal with theoretical concepts that are heterogeneous. In fact, it can be stated that one of the major factors that limit our ability to generalize in ecology is that we seek to generalize across variable entities of interest, such as communities and ecosystems (Grace, 2006).

Composites have, we believe, great potential to facilitate our ability to create models that are empirically meaningful and also of theoretical relevance. Without composites, models that consider substantial complexity and/or seek to address relations among a large number of variables will have the tendency to be highly specific, possessing a separate set of effects from each of a suite of intercorrelated indicators. Such specificity will limit the utility of SEM for effectively addressing the general questions often asked by ecologists. Composites represent something of a compromise, but one that often matches the characteristics of our theoretical constructs reasonably well. In the examples in this paper, using composites allows us to emphasize the interpretation of only 4 pathways from our two main predictor constructs instead of the 10 pathways that would require emphasis without the use of composites.

It should be pointed out that some have argued that because composites have data-specific properties defined by the relationships between predictors and responses, their meaning is vague, their effects less general, and they should be avoided (see discussion in Edwards 2001). Others, ourselves included, feel that the benefits associated with using composites outweigh their limitations. To some degree, the differences in opinion between these two perspectives may depend on the context and the emphasis of the investigator relative to the elements in figure 1. For those who wish to maximize their ability to relate statistical findings to some general theory, composites may provide a very useful device (e.g., Laughlin and Grace, 2006). For other investigators, who either possess theories that make very specific predictions or who do not wish to generalize far from the data, composites may be less valuable.

It should be clear from the examples considered in this paper that the standard latent variable model is not always adequate for representing constructs and their interrelations. A consideration of composite variables contributes balance to the modeling process by providing options and encouraging a thorough evaluation of constructs and the measures at hand. While the need for composite variable specification has been pointed out several times (Heise, 1972; Bollen and Lennox, 1991; MacCallum and Brown, 1993), the inclusion of composites in SEMs remains rare, while the misspecification of models using only $L \rightarrow M$ blocks is common (Jarvis and others, 2003).

At present, much of the literature on composite variables is based on the premise that composites should have two or more effects on other variables (outward directed paths) and that composites with single effects should not be included in models

(MacCallum and Browne, 1993; Edwards, 2001; Jarvis and others, 2003; Williams and others, 2003).

This recommendation has been offered as a necessary, though insufficient, requirement for the identification of composite error variances. In our first example, we found such a model (Model C) to be appropriate and satisfactory; in many cases models of this sort will not be adequate, and those with single effects (e.g., Model E) will be required. We argue that (1) including composites that have only single outward directed paths can be quite valuable, even though the covariances for a set of data can be represented without them, and (2) confining the use of composites to cases where there are multiple outward directed pathways is ill advised because it can lead the researcher into developing and testing models that fail to match the relations in the data.

A key factor that has limited the use of composites has been the aforementioned problem of parameter identification for models with single effects (as well as some models with multiple effects). In this paper, we show that this problem can be resolved by reference to a partially reduced form model (of the type represented by Model C) in which specific effects can be evaluated, in conjunction with the specification of a scale for the composite. Further, we recommend that, for models that incorporate composites, reduced form models such as Model C and single-effect models such as Model D should be routinely examined as part of the evaluation process. When this approach is taken, a more complete consideration of construct representation can be achieved.

As illustrated through our considerations of Example 2, representing endogenous constructs by using composites poses additional theoretical and practical issues. Models of the most general type, which posit direct effects of exogenous variables on endogenous composites, while of interest, may be difficult to solve and also to interpret. Models in which effects on endogenous constructs are modeled through influences on their dimensions prove to be more tractable, as well as more generally interpretable. Again, as a general strategy, comparing results with those from models with fewer restrictions contributes to the evaluation process.

Modern SEM using latent variables has largely developed through applications to the social and economic sciences, though it is now being widely used in a broad range of disciplines. Part of the emphasis on common factor-type models (multi-indicator, latent variable models) certainly comes from the availability of redundant measures, such as batteries of questionnaire items or exam questions that can be used to estimate underlying causes. Even in these disciplines, it seems that data and constructs

are such that composites could be widely used to represent relationships more properly. In the natural sciences, redundant measures are less common, a fact that has contributed to an emphasis on path analysis using only observed variables and encouraged little use of latent variables. We expect that multi-indicator latent variables of the L→M block type will find wide usage in the natural sciences over time. However, we anticipate that a significant need will develop for considerations of composite variables to represent collections of effects. Composites, when used with care, provide a useful way of summarizing groups of effects that match our theories about construct relations.

Acknowledgments

We wish to thank K. Verheyen, G.R. Guntenspergen, B. Biesbrouck, and M. Hermy for use of their data (with permission) in Example 1. We also thank J.S. Lichtenberg, S.L. King, and S. Walls for use of their unpublished data in Example 2. We are grateful to Timothy Gregoire and three anonymous reviewers for constructive comments that aided with the revision of the manuscript. Thanks also to Bill Shipley for helpful comments on an earlier version of this manuscript and to Connie Herndon, Victoria Jenkins, and Darren Johnson for additional reviews.

References Cited

- Belovsky, G.E., Botkin, D.B., Crowl, T.A., Cummins, K.W., Franklin, J.F., Hunter, M.L., Joern, A., Lindemayer, D.B., MacMahon, J.A., Margules, C.R., and Scott, M.J., 2004, Ten suggestions to strengthen the science of ecology: *Bioscience*, v. 54, p. 345—351.
- Blalock, H.M., 1964, Causal inferences in nonexperimental research: Chapel Hill, University of North Carolina Press.
- Bollen, K.A., 1984, Multiple indicators: internal consistency or no necessary relationship. *Quality and Quantity*, v. 18, p. 377—385.
- Bollen, K.A., 1989, Structural equations with latent Variables: New York, John Wiley & Sons.
- Bollen, K.A., and Lennox, R., 1991, Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin*, v. 110, p. 305—314.
- Bollen, K.A., and Ting, K., 2000, A tetrad test for causal indicators: *Psychological Methods*, v. 5, p. 3—22.
- Diamantopoulous, A., and Winklhofer, H.M., 2001, Index construction with formative indicators: an alternative to scale development: *Journal of Marketing Research*, v. 38, p. 269—277.
- Edwards, J.R., 2001, Multidimensional constructs in organizational behavior research: an integrative analytical framework: *Organizational Research Methods*, v. 4, p. 144—192.
- Fayers, P.M., and Hand, D.J., 2002, Causal variables, indicator variables and measurement scales: an example from quality of life. *Journal of the Royal Statistical Society: Series A*, v. 165, p. 233—261.
- Grace, J.B., 2003, Examining the relationship between environmental variables and ordination axes using latent variables and structural equation modeling. chap. 7 (Pages 171—193), in Pugesek, B.H., Tomer, A., and von Eye, A., eds., *Structural equation modeling*, Cambridge, UK, Cambridge University Press, p. 171—193.
- Grace, J.B., 2006, Structural equation modeling and natural systems. Cambridge, UK, Cambridge University Press.
- Grace, J.B., and Bollen, K.A., 2005, Interpreting the results from multiple regression and structural equation models: *Bulletin of the Ecological Society of America*, v. 86, p. 283—295.
- Grace, J.B., and Bollen, K.A., in press, Representing general theoretical concepts in structural equation models: the role of composite variables: *Environmental and Ecological Statistics*, v. 0, p. 000—000.
- Heise, D.R., 1972, Employing nominal variables, induced variables, and block variables in path analyses: *Sociological Methods & Research*, v. 1, p. 147—173.
- Hipp, J., Bauer, D.J., and Bollen, K.A., 2005, Conducting tetrad tests of model fit and contrasts of tetrad-nested models: a new SAS macro: *Structural Equation Modeling*, v. 12, p. 76—93.
- Jarvis, C.B., MacKenzie, S.B., and Podsakoff, P.M., 2003, A critical review of construct indicators and measurement model misspecification in marketing and consumer research: *Journal of Consumer Research*, v. 30, p. 199—218.

- Johnson, M.L., Huggins, D.G., and deNoyelles, F., Jr., 1991, Ecosystem modeling with LISREL: a new approach for measuring direct and indirect effects: *Ecological Applications*, v. 1, p. 383—398.
- Laughlin, D.C. and Grace, J.B., 2006, A multivariate model of plant species richness in forested systems: old-growth montane forests with a long history of fire. *Oikos*, v. 114, p. 60—70.
- Levins, R., 1968, *Evolution in changing environments*, Princeton, Princeton University Press.
- Lichtenberg, J.S., King, S.L., Grace, J.B., and Walls, S., 2006, Habitat associations of chorusing anurans in the lower Mississippi river alluvial valley: *Wetlands*, v. 26, p. 000—000.
- MacCallum, R.C., and Browne, M.W., 1993, The use of causal indicators in covariance structure models: some practical issues: *Psychological Bulletin*, v. 114, p. 533—541.
- Mitchell, R.J., 1992, Testing evolutionary and ecological hypotheses using path analysis and structural equation modelling: *Functional Ecology*, v. 6, p. 123—129.
- Muthén, L.K., and Muthén, B.O., 2004, *Mplus user's guide*. (3rd ed.): Los Angeles, Muthén and Muthén.
- Pugesek, B.H., and Tomer, A., 1996, The Bumpus house sparrow data: a reanalysis using structural equation models: *Evolutionary Ecology*, v. 10, p. 387—404.
- Pugesek, B.H., Tomer, A., and von Eye, A., 2003, *Structural equation modeling*: Cambridge: UK, Cambridge University Press.
- Shipley, B., 2000. *Cause and correlation in biology*: Cambridge, UK, Cambridge University Press.
- Spirtes, P., Glymour, C., and Scheines, R., 1993, *Causation, prediction, and search*: New York, Springer-Verlag.
- Verheyen, K., Guntenspergen, G.R., Biesbrouck, B., and Hermy, M., 2003, An integrated analysis of the effects of past land use on forest herb colonization at the landscape scale: *Journal of Ecology*, v. 91, p. 731—742.
- Weiher, E., Forbes, S., Schauwecker, T., and Grace, J.B., 2004, Multivariate control of plant species richness in a blackland prairie: *Oikos*, v. 106, p. 151—157.
- Williams, L.J., Edwards, J.R., and Vandenberg, R.J., 2003, Recent advances in causal modeling methods for organizational and management research. *Journal of Management*, v. 29, p. 903—936.
- Wright, S., 1921, Correlation and causation: *Journal of Agricultural Research*, v. 10, p. 557—585.
- Wright, S., 1968. *Evolution and the genetics of populations*, Vol. 1: Genetic and Biometric Foundations, Chicago, University of Chicago Press.